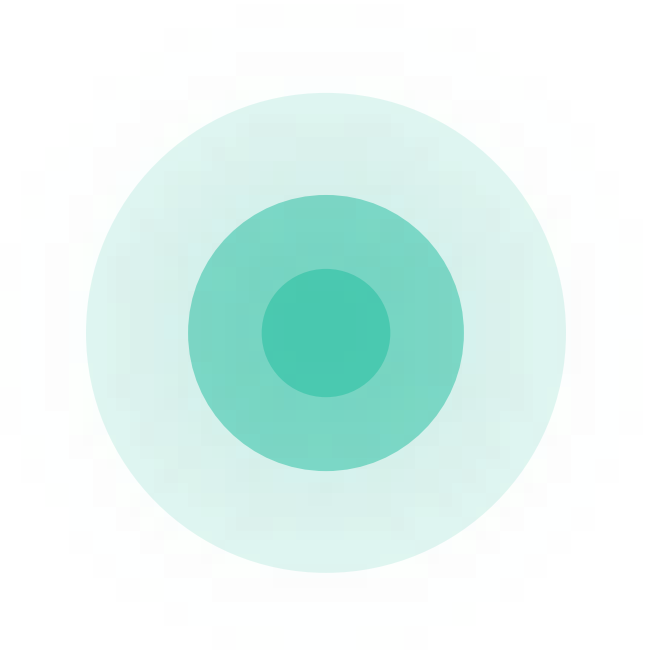


Microsoft Azure User Group Poland

21.06.2017

Machine Learning w Microsoft Azure

Mateusz Grzyb
konsultant technologiczny Microsoft Polska
mateuszgrzyb.pl

O czym będzie ta prezentacja?

Plan prezentacji

1. Machine Learning.
2. Machine Learning w Microsoft Azure.
3. Microsoft Azure Machine Learning Studio.
4. Klasyfikacja pasażerów Titanica z użyciem Microsoft Azure Machine Learning Studio.
5. Pytania do Was.
6. Pytania do mnie.

Machine Learning

Machine Learning

Interdyscyplinarna (matematyka, robotyka, informatyka) dziedzina naukowa wchodząca w skład nauk zajmujących się SI.

Celem ML jest:

- Zastosowanie w SI do zbudowania automatycznego systemu gromadzącego doświadczenie (na podstawie danych uczących) i nabywanie na tej podstawie nowej wiedzy.
- Rozwiązywanie problemów w sposób automatyczny.

Typy problemów jakie rozwiązuje ML

Typy problemów jakie rozwiązuje ML

- Klasyfikacja (dwuklasowa i wieloklasowa), np. dobry/zły kredytobiorca.
- Regresja, np. ile warte jest dane mieszkanie?
- Klasteryzacja, np. dzielenie klientów banku na klastry w oparciu o historię kredytową i produkty z których korzystali.
- Inne – analiza sentymentu, wykrywanie anomalii, algorytmy rekomendacyjne.

Mierzenie skuteczności działania algorytmów klasyfikacyjnych

Mierzenie skuteczności działania algorytmów klasyfikacyjnych?

		Predykcja	
		0	1
Rzeczywiste wartości	0	TN	FP (błąd I rodzaju)
	1	FN (błąd II rodzaju)	TP

- Dokładność (ang. *Accuracy*) – ACC
  ACC = (TP+TN)/(P+N)
- Precyzja (ang. *Precision*) – PPV
  PPV = TP/(TP+FP)
- Czułość (ang. *sensitivity*) – TPR
  TPR = TP/P = TP/(TP+FN)

Mierzenie skuteczności działania algorytmów klasyfikacyjnych?

		Predykcja	
		0	1
Rzeczywiste wartości	0	TN	FP (błąd I rodzaju)
	1	FN (błąd II rodzaju)	TP

- Dokładność (ang. *Accuracy*) – ACC
  ACC = (TP+TN)/(P+N)
- Precyzja (ang. *Precision*) – PPV
  PPV = TP/(TP+FP)
- Czułość (ang. *sensitivity*) – TPR
  TPR = TP/P = TP/(TP+FN)

Zastosowanie Machine Learning

allegro

Machine Learning w Microsoft Azure

Cortana Intelligence Suite

Intelligent Apps require Intelligent Solutions

Machine Learning w Microsoft Azure

R/Python with HDInsight

R, Spark i MapReduce do wykonywania obliczeń rozproszonych.

MicrosoftML with SQL Server

Dedykowana biblioteka Machine Learning dla R w SQL Server.

Machine Learning Studio

Usługa Azure Machine Learning.

Machine Learning w Microsoft Azure

Azure Data Science Virtual Machine

Dedykowana wirtualna maszyna przygotowana z myślą o Data Scientist. Zawiera preinstalowane oprogramowanie niezbędne do analizy i wizualizacji danych:

- Microsoft R Server Developer Edition
- Python Anaconda
- Julia Pro
- Power BI Desktop
- SQL Server 2016 Developer edition (R included)
- CNTK 2.0
- i wiele więcej...

Machine Learning w Microsoft Azure

Nowość: Azure Data Science Virtual Machine + lokalna sesja R

```
# dedykowana biblioteka
library(devtools)
devtools::install_github("Azure/AzureDSVM")
library("AzureDSVM")

# powoływanie pojedynczej maszyny do życia (ok. 4 min)
ldsvm <- deployDSVM(context, resource.group="example",
            location="southeastasia", size="Standard_D4_v2",
            os="Ubuntu", hostname="mydsvm",
            username="myname", pubkey="pubkey")
```

Machine Learning w Microsoft Azure

Nowość: Azure Data Science Virtual Machine + lokalna sesja R

```
# klaster złożony z 5 maszyn
cluster <-deployDSVMCluster(context, resource.group=RG, location="southeastasia",
hostnames="mydsvm", usernames="myname", pubkeys="pubkey", count=5)

# context – obiekt utworzony przez funkcję  createAzureContext(), zawierającą poświadczenia w
postaci: id tenanta, id klienta, etc.
```

Machine Learning w Microsoft Azure

Nowość: Azure Data Science Virtual Machine + lokalna sesja R

```
# wykonywanie zdalne lokalnego skryptu z biblioteki Microsoft RevoScaleR na maszynie wirtualnej
code <- paste(„  library(scales)", "df <- scale(iris[, -5])", "rxExec(kmeans, x=df, centers=2)", sep=";")
tmpf1 <- tempfile(paste0("AzureDSVM_experiment_01_"))
file.create(tmpf1)
writeLines(code, tmpf1)

executeScript(context, resource.group=RG, machines=cluster$hostname[1],
        remote=cluster$fqdn[1], user=unique(cluster$username), script=tmpf1,
        master=cluster$fqdn[1], slaves=cluster$fqdn[1],
        compute.context="localParallel/clusterParallel")
```

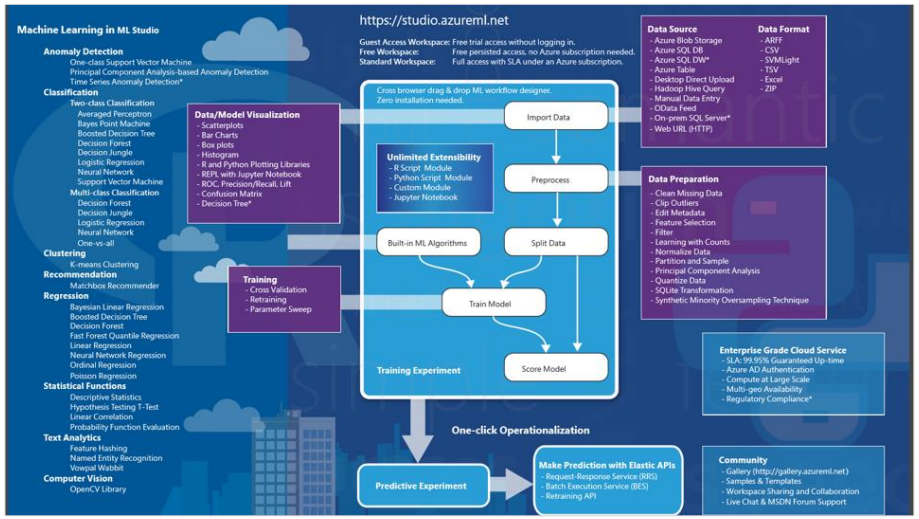
Microsoft Azure
Machine Learning Studio

Microsoft Azure Machine Learning Studio

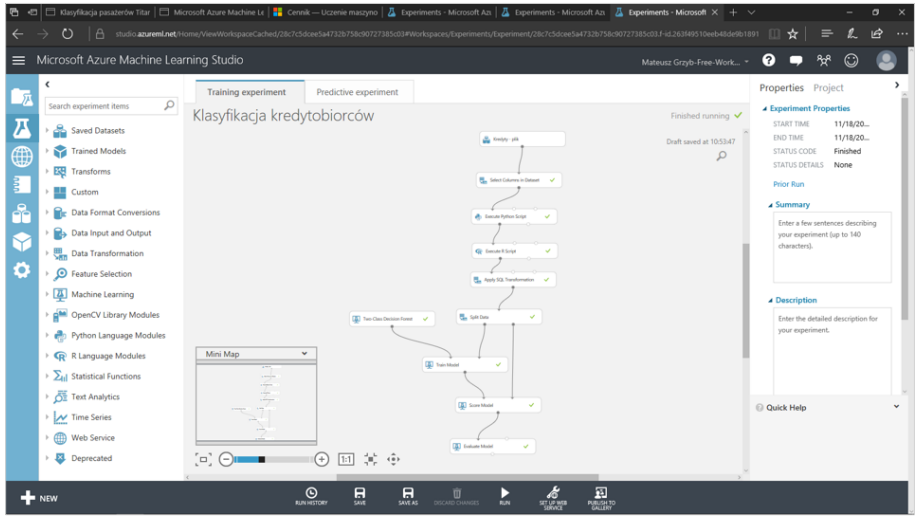
Oprogramowanie typu *drag and drop*, zawierające:

- wbudowany zestaw komponentów do obróbki i wizualizacji danych,
- zestaw dedykowanych algorytmów uczenia maszynowego,
- mechanizmy pozwalające wczytywać dane składowane OnPremises,
- funkcjonalność generowania web serwisu, który pozwala na udostępnianie modelu predykcyjnego „na zewnątrz”.

Microsoft Azure Machine Learning Studio



Azure ML Studio – przegląd funkcjonalności



Przykład działania Azure ML Studio

Microsoft Azure ML Studio - wady i zalety

Prostota

Szybkość implementacji rozwiązania

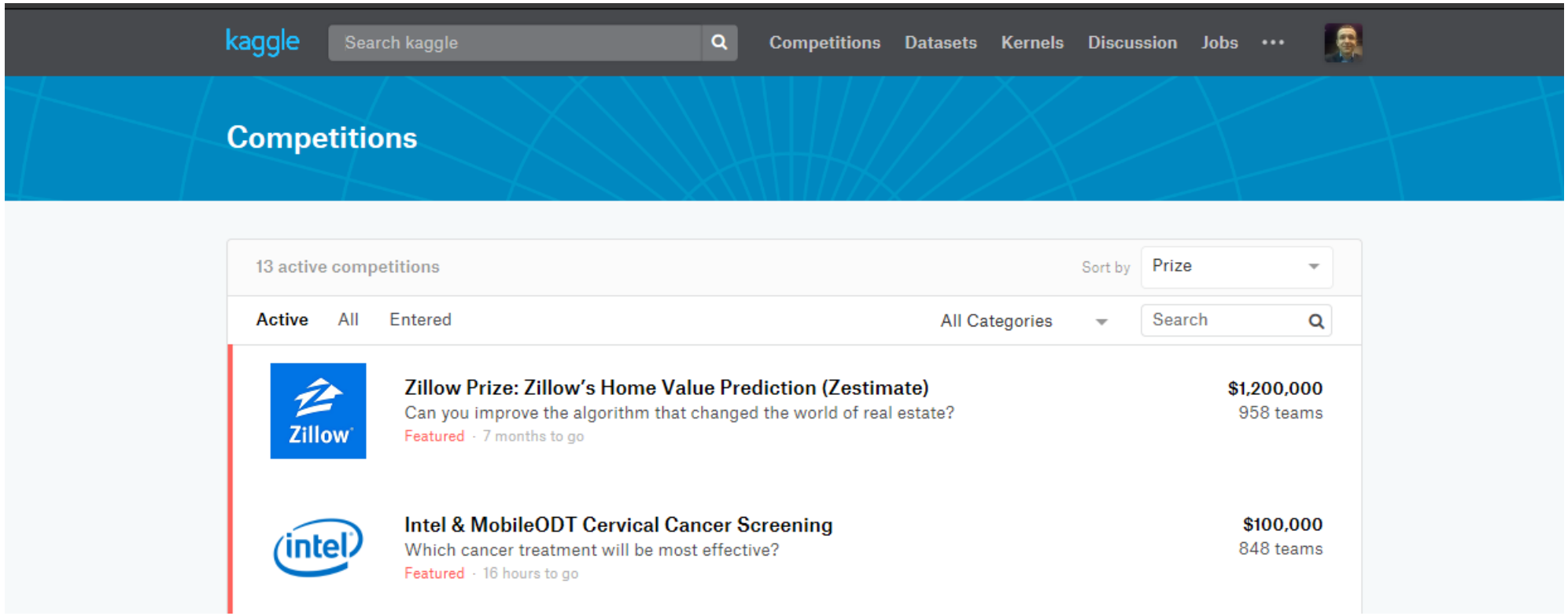
Brak wglądu „pod maskę” rozwiązania

Klasyfikacja pasażerów Titanica

Czym jest Kaggle?

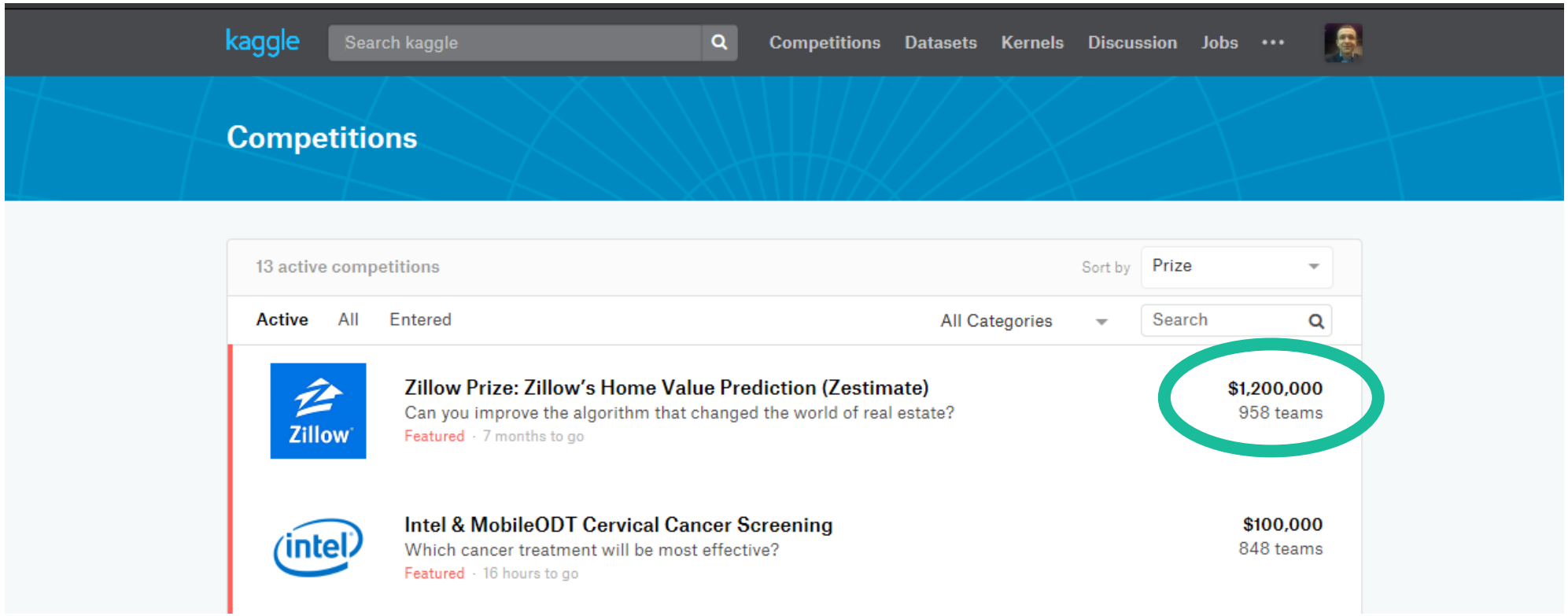
Czym jest Kaggle?

Platforma konkursowa kojarząca ze sobą Data Scientistów i firmy poszukujące rozwiązania na trapiące je problemy.



Czym jest Kaggle?

Platforma konkursowa kojarząca ze sobą Data Scientistów i firmy poszukujące rozwiązania na trapiące je problemy.



Data Science Bowl 2017

Can you improve lung cancer detection?

$1,000,000 · 1,972 teams · 2 months ago

Overview

In the United States, lung cancer strikes 225,000 people every year, and accounts for $12 billion in health care costs. Early detection is critical to give patients the best chance at recovery and survival.

One year ago, the office of the U.S. Vice President spearheaded a bold new initiative, the Cancer Moonshot, to make a decade's worth of progress in cancer prevention, diagnosis, and treatment in just 5 years.

In 2017, the Data Science Bowl will be a critical milestone in support of the Cancer Moonshot by convening the data science and medical communities to develop lung cancer detection algorithms.

Using a data set of thousands of high-resolution lung scans provided by the National Cancer Institute, participants will develop algorithms that accurately determine when lesions in the lungs are cancerous. This will dramatically reduce the false positive rate that plagues the current detection technology, get patients earlier access to life-saving interventions, and give radiologists more time to spend with their patients.

This year, the Data Science Bowl will award $1 million in prizes to those who observe the right patterns, ask the right questions, and in turn, create unprecedented impact around cancer screening care and prevention. The funds for the prize purse will be provided by the Laura and John Arnold Foundation.

Visit DataScienceBowl.com to:

Data Science Bowl 2017

Can you improve lung cancer detection?

$1,000,000 · 1,97 teams · 2 months ago

Overview | Data | Kernels | Discussion | Leaderboard | More | Submit Predictions

Overview

- Description
- Evaluation
- Prizes
- About
- Engagement Contest
- Resources
- Timeline
- Tutorial

In the United States, lung cancer strikes 225,000 people every year, and accounts for $12 billion in health care costs. Early detection is critical to give patients the best chance at recovery and survival.

One year ago, the office of the U.S. Vice President spearheaded a bold new initiative, the Cancer Moonshot, to make a decade's worth of progress in cancer prevention, diagnosis, and treatment in just 5 years.

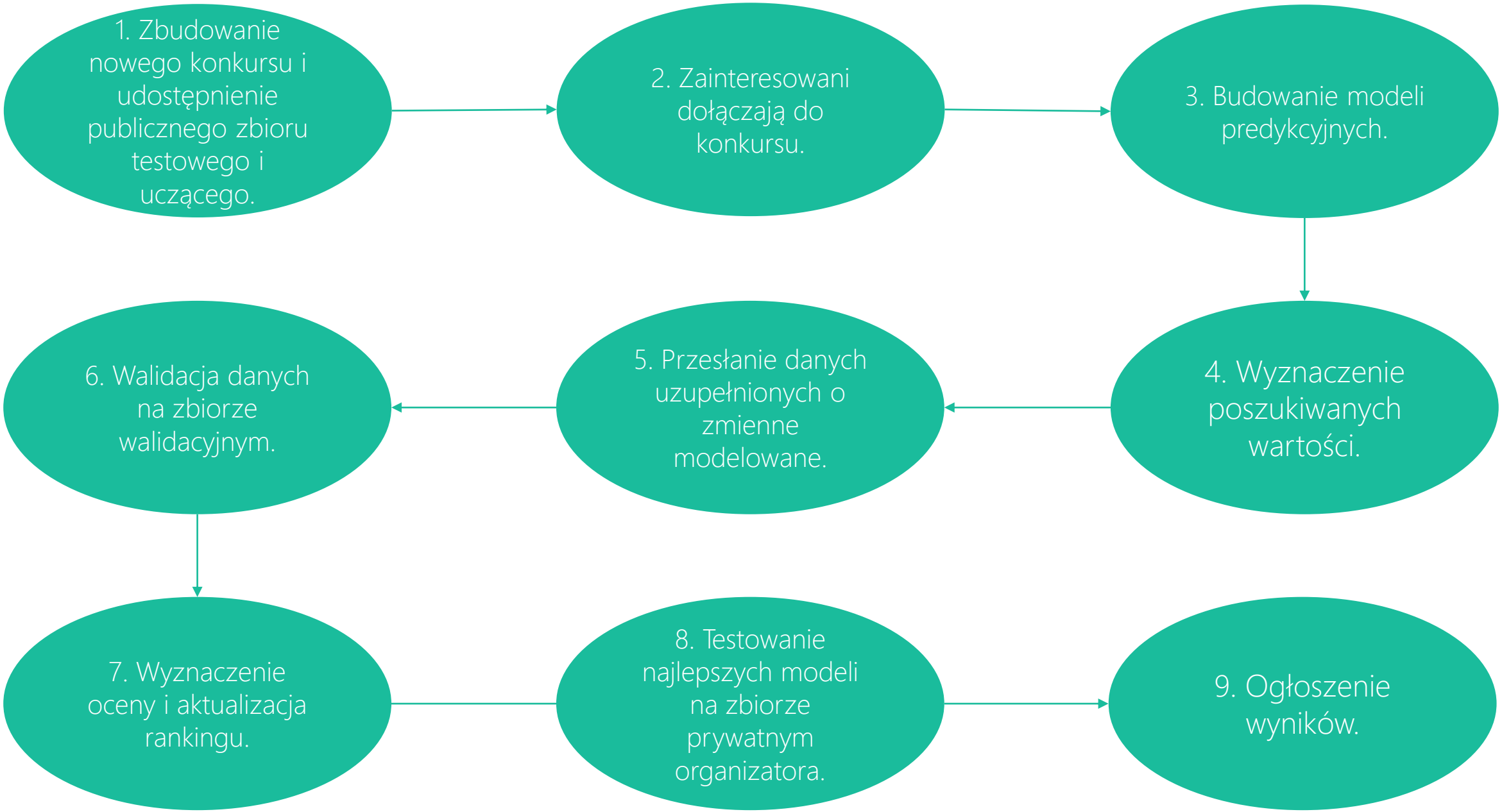
In 2017, the Data Science Bowl will be a critical milestone in support of the Cancer Moonshot by convening the data science and medical communities to develop lung cancer detection algorithms.

Using a data set of thousands of high-resolution lung scans provided by the National Cancer Institute, participants will develop algorithms that accurately determine when lesions in the lungs are cancerous. This will dramatically reduce the false positive rate that plagues the current detection technology, get patients earlier access to life-saving interventions, and give radiologists more time to spend with their patients.

This year, the Data Science Bowl will award $1 million in prizes to those who observe the right patterns, ask the right questions, and in turn, create unprecedented impact around cancer screening care and prevention. The funds for the prize purse will be provided by the Laura and John Arnold Foundation.

Visit DataScienceBowl.com to:

Jak działają konkursy na Kaggle?


1. Zbudowanie nowego konkursu i udostępnienie publicznego zbioru testowego i uczącego.
2. Zainteresowani dołączają do konkursu.
3. Budowanie modeli predykcyjnych.
4. Wyznaczenie poszukiwanych wartości.
5. Przesłanie danych uzupełnionych o zmienne modelowane.
6. Walidacja danych na zbiorze walidacyjnym.
7. Wyznaczenie oceny i aktualizacja rankingu.
8. Testowanie najlepszych modeli na zbiorze prywatnym organizatora.
9. Ogłoszenie wyników.

Klasyfikacja pasażerów Titanica

Klasyfikacja pasażerów Titanica

Jeden z najpopularniejszych konkursów na Kaggle: 7257 zespołów.
Ranking zawiera wyniki nie starsze niż 2 miesiące.
Wynik jest mierzony poprzez Accuracy: ACC = (TP+TN)/(P+N)
Brak nagrody głównej.

Cel konkursu:
Na podstawie zbioru uczącego (891 obserwacji), sklasyfikować 418 pasażerów płynących na Titanicu, tzn. przewidzieć czy danej osobie udało się przeżyć katastrofę.

Cel konkursu

Zbudowanie modelu, który w oparciu o wybrane spośród 10 predyktorów (atrybutów predykcyjnych) pozwoli nam wyznaczyć wartość zmiennej wynikowej.

Atrybutem, którego wartość będziemy przewidywać jest „Survival", który mówi nam czy dany pasażer przeżył katastrofę, czy też nie. Przyjmuje on wartości [0,1], gdzie 0=Nie, 1=Tak.

Opis zbioru uczącego

By wyznaczyć wartość „Survival”, można użyć zmiennych:

- Pclass – klasa | (1 = pierwsza; 2 = druga; 3 = trzecia).
- Name – imię i nazwisko pasażera.
- Sex – płeć pasażera.
- Sibsp – liczba małżonków, lub rodzeństwa na pokładzie.
- Parch – liczba rodziców, lub dzieci na pokładzie.
- Ticket – numer biletu.
- Fare – opłata za bilet.
- Cabin – kabina.
- Embarked – port startowy (C = Cherbourg; Q = Queenstown; S = Southampton).

Mój cel?

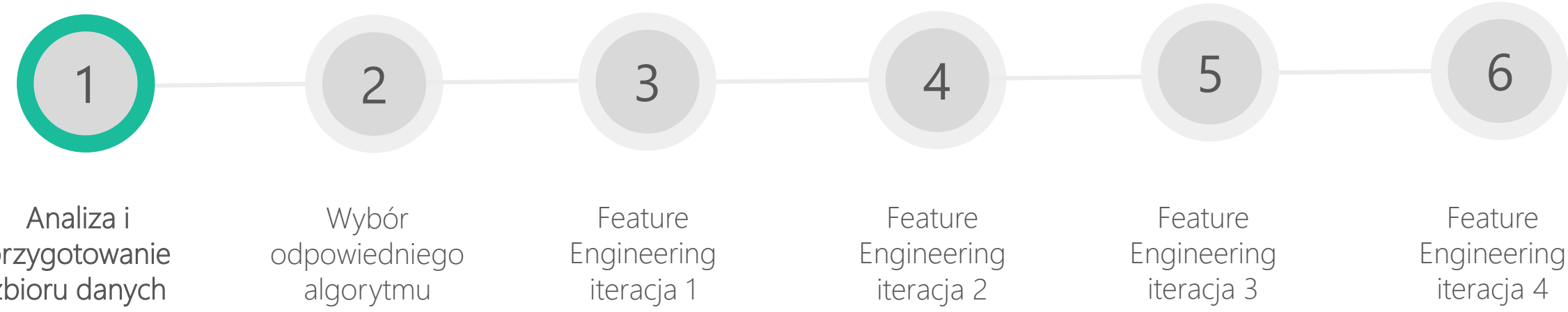
ACC > 0.77990

Proces budowy modelu predykcyjnego

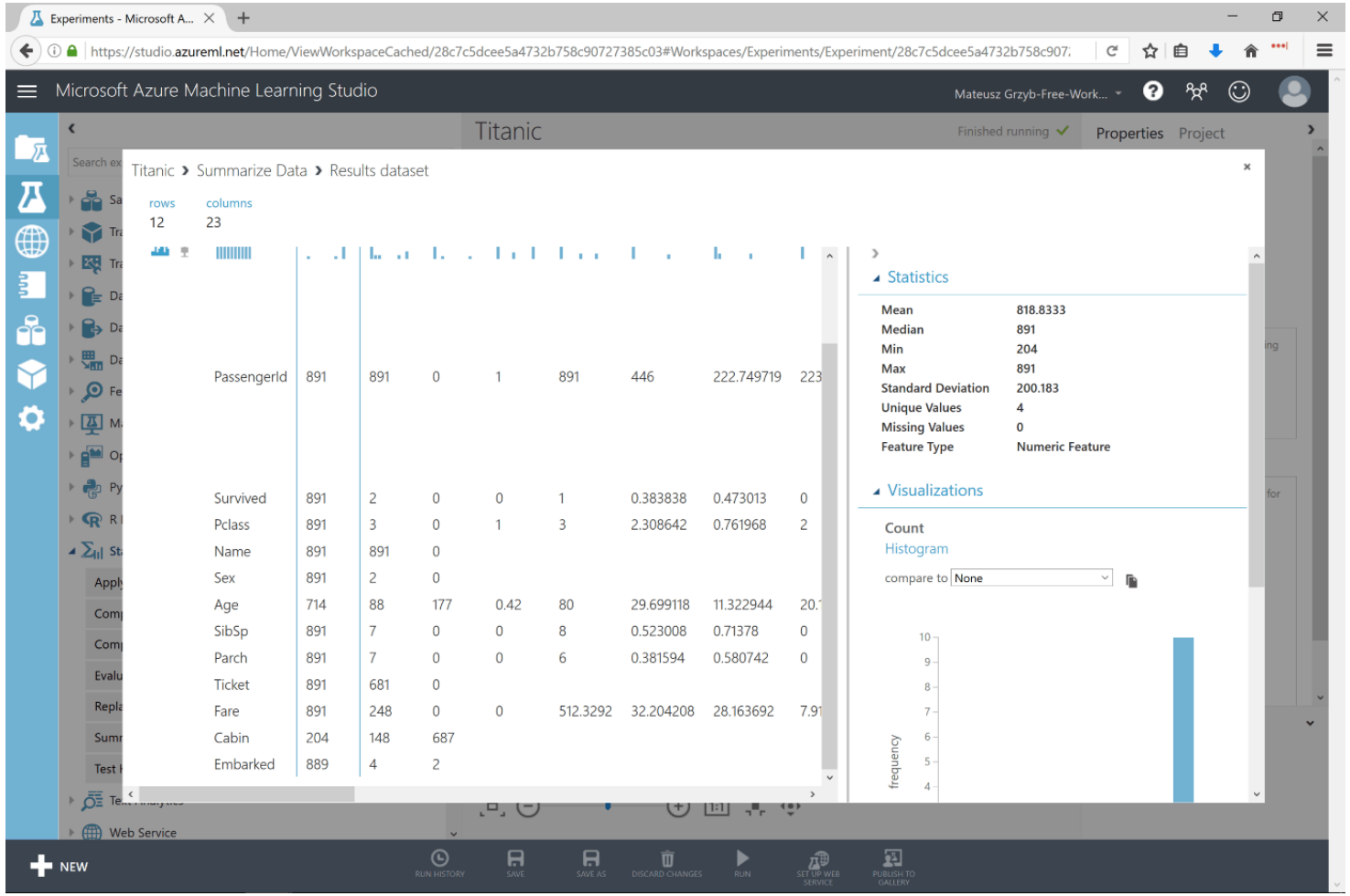
Proces budowy modelu predykcyjnego

1. Analiza i przygotowanie zbioru danych
2. Wybór odpowiedniego algorytmu
3. Feature Engineering iteracja 1
4. Feature Engineering iteracja 2
5. Feature Engineering iteracja 3
6. Feature Engineering iteracja 4

Proces budowy modelu predykcyjnego



Analiza i przygotowanie zbioru danych



Analiza i przygotowanie zbioru danych

Pierwsze utrudnienia:

- Kolumny: „Survived" i „Pclass" są błędnie interpretowane przez Azure ML jako zmienne numeryczne. Są to jednak bez wątpienia zmienne kategoryczne.
- Kolumny: „Embarked" i „Sex" są błędnie interpretowane przez Azure ML jako zmienne tekstowe. Z punktu widzenia badacza danych są to jednak zmienne kategoryczne.
- Istnieje kilkadziesiąt brakujących wartości dla atrybutów: „Age", „Cabin" (ponad 700 braków, rekompensuje je „Pclass"), „Embarked".

Edycja metadanych

Experiments - Microsoft A... | Experiments - Microsoft A...

https://studio.azureml.net/Home/ViewWorkspaceCached/28c7c5dcee5a4732b758c90727385c03#Workspaces/Experiments/Experiment/28c7c5dcee5a4732b758c907

Microsoft Azure Machine Learning Studio

Mateusz Grzyb-Free-Work...

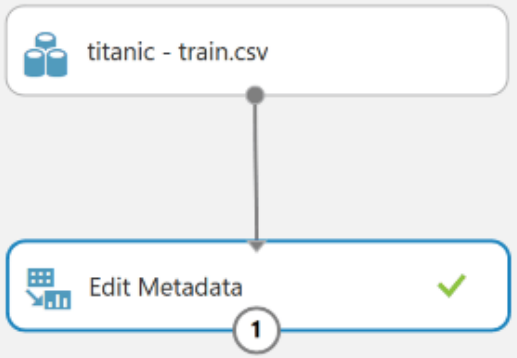
Search experiment items

- Saved Datasets
- Trained Models
- Transforms
- Data Format Conversions
- Data Input and Output
- Data Transformation
  - Filter
  - Learning with Counts
  - Manipulation
    - Add Columns
    - Add Rows
    - Apply SQL Transformation
    - Clean Missing Data
    - Convert to Indicator Values
    - Edit Metadata
    - Group Categorical Values
    - Join Data
    - Remove Duplicate Rows
    - Select Columns in Dataset

Titanic

Finished running

Draft saved at 17:04:31



Properties | Project

Edit Metadata

Column

Selected columns:
Column names:
Survived,Pclass,Embarked,Sex

Launch column selector

Data type: Unchanged

Categorical: Make categorical

Fields: Unchanged

New column names

START TIME	1/24/2017 5:03:28 PM
END TIME	1/24/2017 5:03:30 PM
ELAPSED TIME	0:00:02.291
STATUS CODE	Finished
STATUS DETAILS	None

View output log

Quick Help

Uzupełnienie brakujących wartości

Experiments - Microsoft A... | Experiments - Microsoft A...

https://studio.azureml.net/Home/ViewWorkspaceCached/28c7c5dcee5a4732b758c90727385c03#Workspaces/Experiments/Experiment/28c7c5dcee5a4732b758c907

Microsoft Azure Machine Learning Studio

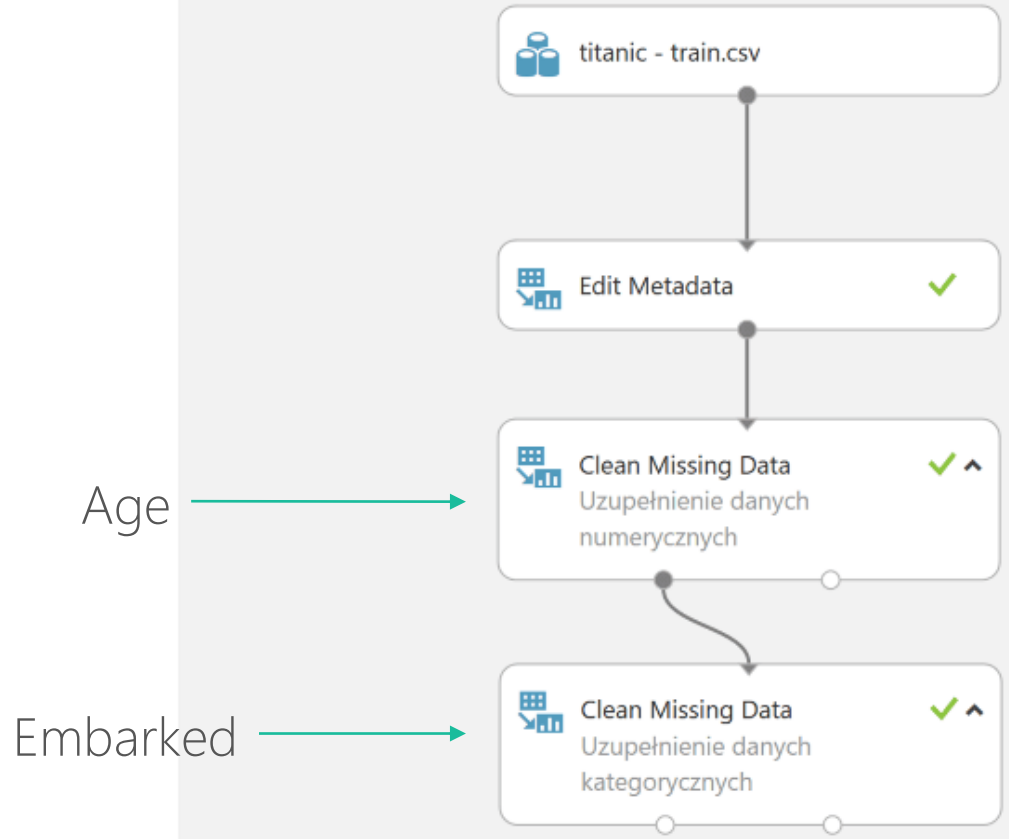
Mateusz Grzyb-Free-Work...

Search experiment items

- Saved Datasets
  - My Datasets
  - Samples
- Trained Models
- Transforms
- Data Format Conversions
- Data Input and Output
- Data Transformation
- Feature Selection
- Machine Learning
- OpenCV Library Modules
- Python Language Modules
- R Language Modules
- Statistical Functions
- Text Analytics
- Web Service
- Deprecated

Titanic

Finished running



Age →

Embarked →

Properties | Project

Experiment Properties

START TIME	1/24/2017 5:21:28 PM
END TIME	1/24/2017 5:21:43 PM
STATUS CODE	Finished
STATUS DETAILS	None

Prior Run

Summary

Enter a few sentences describing your experiment (up to 140 characters).

Description

Enter the detailed description for your experiment.

Quick Help

Normalizacja danych numerycznych

Analiza i przygotowanie zbioru danych

Po co normalizować dane numeryczne?

Podczas uczenia maszynowego zmienne numeryczne o większych wartościach mogą być postrzegane przez algorytm jako ważniejsze.

Przykład: Dana pasażerka o imieniu *Sandstrom, Miss. Marguerite Rut* szczęśliwie przeżyła katastrofę Titanica. Z danych jasno wynika, że zapłacił za bilet 16.7, a w chwili podróży miała jedynie 4 lata.

Algorytm uzna, że zdecydowanie większy wpływ na to, że udało mu się ocaleć miała cena biletu, co oczywiście nie musi być prawdą.

Przycięcie odstających wartości

Titanic

In draft

Properties Project

Titanic › Clean Missing Data › Cleaned dataset

rows 891

columns 12

view as

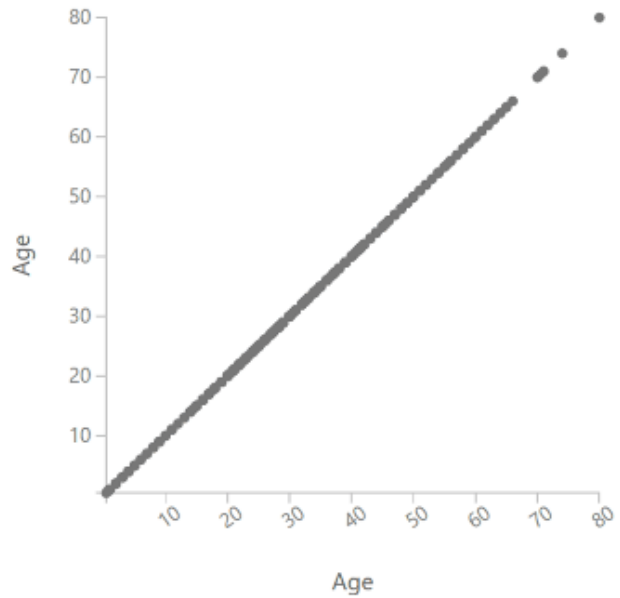
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	T
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	P
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	1
5	0	3	Allen, Mr. William Henry	male	35	0	0	3
6	0	3	Moran, Mr. James	male	28.204041	0	0	3
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	1
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	3
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	3

Visualizations

Age

ScatterPlot

compare to Age



Age log scale

Age log scale

Group Categorical Values

Titanic

In draft

Properties Project

Titanic › Clip Values › Results dataset

rows 891

columns 12

view as

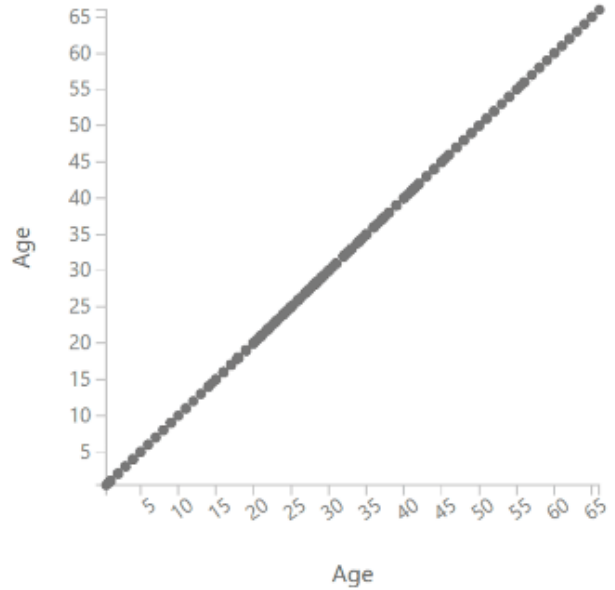
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	T
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	P
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	1
5	0	3	Allen, Mr. William Henry	male	35	0	0	3
6	0	3	Moran, Mr. James	male	28.204041	0	0	3
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	1
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	3
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	3

Visualizations

Age

ScatterPlot

compare to Age

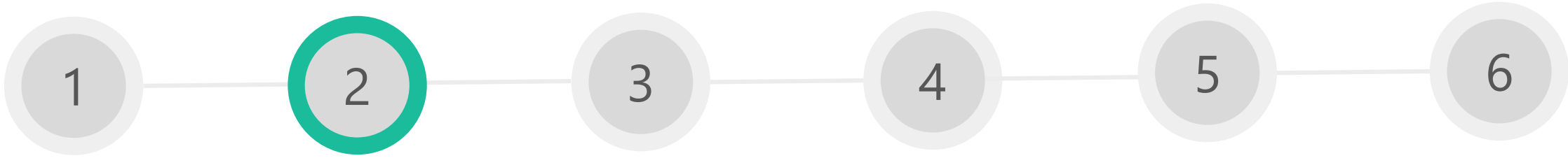


Age log scale

Age log scale

Group Categorical Values

Proces budowy modelu predykcyjnego



1. Analiza i przygotowanie zbioru danych
2. **Wybór odpowiedniego algorytmu**
3. Feature Engineering iteracja 1
4. Feature Engineering iteracja 2
5. Feature Engineering iteracja 3
6. Feature Engineering iteracja 4

Microsoft Azure Machine Learning: Algorithm Cheat Sheet

This cheat sheet helps you choose the best Azure Machine Learning Studio algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the question you're trying to answer.

START

Finding unusual data points

Discovering structure

Three or more

Predicting categories

Predicting values

Two

ANOMALY DETECTION

- **One-class SVM** — >100 features, aggressive boundary
- **PCA-based anomaly detection** — Fast training

CLUSTERING

- **K-means**

MULTI-CLASS CLASSIFICATION

- Fast training, linear model — **Multiclass logistic regression**
- Accuracy, long training times — **Multiclass neural network**
- Accuracy, fast training — **Multiclass decision forest**
- Accuracy, small memory footprint — **Multiclass decision jungle**
- Depends on the two-class classifier, see notes below — **One-v-all multiclass**

REGRESSION

- **Ordinal regression** — Data in rank ordered categories
- **Poisson regression** — Predicting event counts
- **Fast forest quantile regression** — Predicting a distribution
- **Linear regression** — Fast training, linear model
- **Bayesian linear regression** — Linear model, small data sets
- **Neural network regression** — Accuracy, long training time
- **Decision forest regression** — Accuracy, fast training
- **Boosted decision tree regression** — Accuracy, fast training, large memory footprint

TWO-CLASS CLASSIFICATION

- **Two-class SVM** — >100 features, linear model
- **Two-class averaged perceptron** — Fast training, linear model
- **Two-class logistic regression** — Fast training, linear model
- **Two-class Bayes point machine** — Fast training, linear model

- Accuracy, fast training — **Two-class decision forest**
- Accuracy, fast training, large memory footprint — **Two-class boosted decision tree**
- Accuracy, small memory footprint — **Two-class decision jungle**
- >100 features — **Two-class locally deep SVM**
- Accuracy, long training times — **Two-class neural network**

© 2015 Microsoft Corporation. All rights reserved.

Created by the Azure Machine Learning Team

Email: AzurePoster@microsoft.com

Download this poster: http://aka.ms/MLCheatSheet

Microsoft

Experiments - Microsoft A...

https://studio.azureml.net/Home/ViewWorkspaceCached/28c7c5dcee5a4732b758c90727385c03#Workspaces/Experiments/Experiment/28c7c5dcee5a4732b758c907

Microsoft Azure Machine Learning Studio

Mateusz Grzyb-Free-Work...

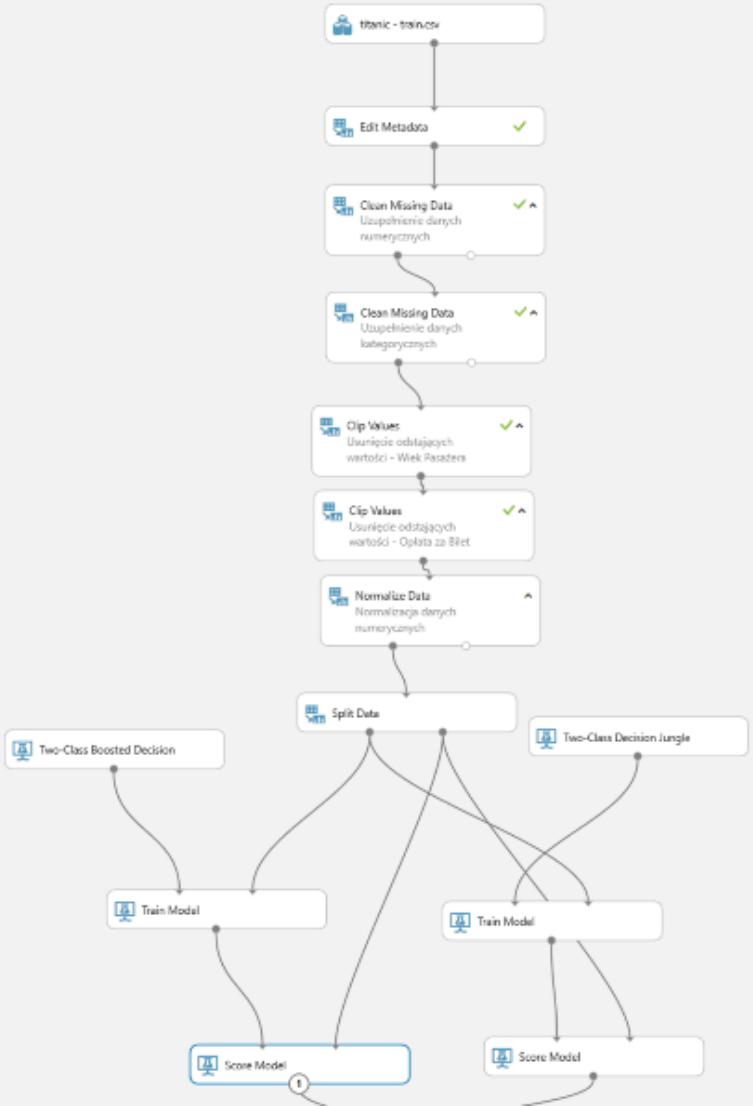
evalu

- Machine Learning
  - Evaluate
    - Cross Validate Model
    - Evaluate Model
    - Evaluate Recommender
- Statistical Functions
  - Evaluate Probability Function

Titanic

In draft

Draft saved at 18:44:04



Properties Project

Score Model

Append score column...

Quick Help

Score a trained classification or regression

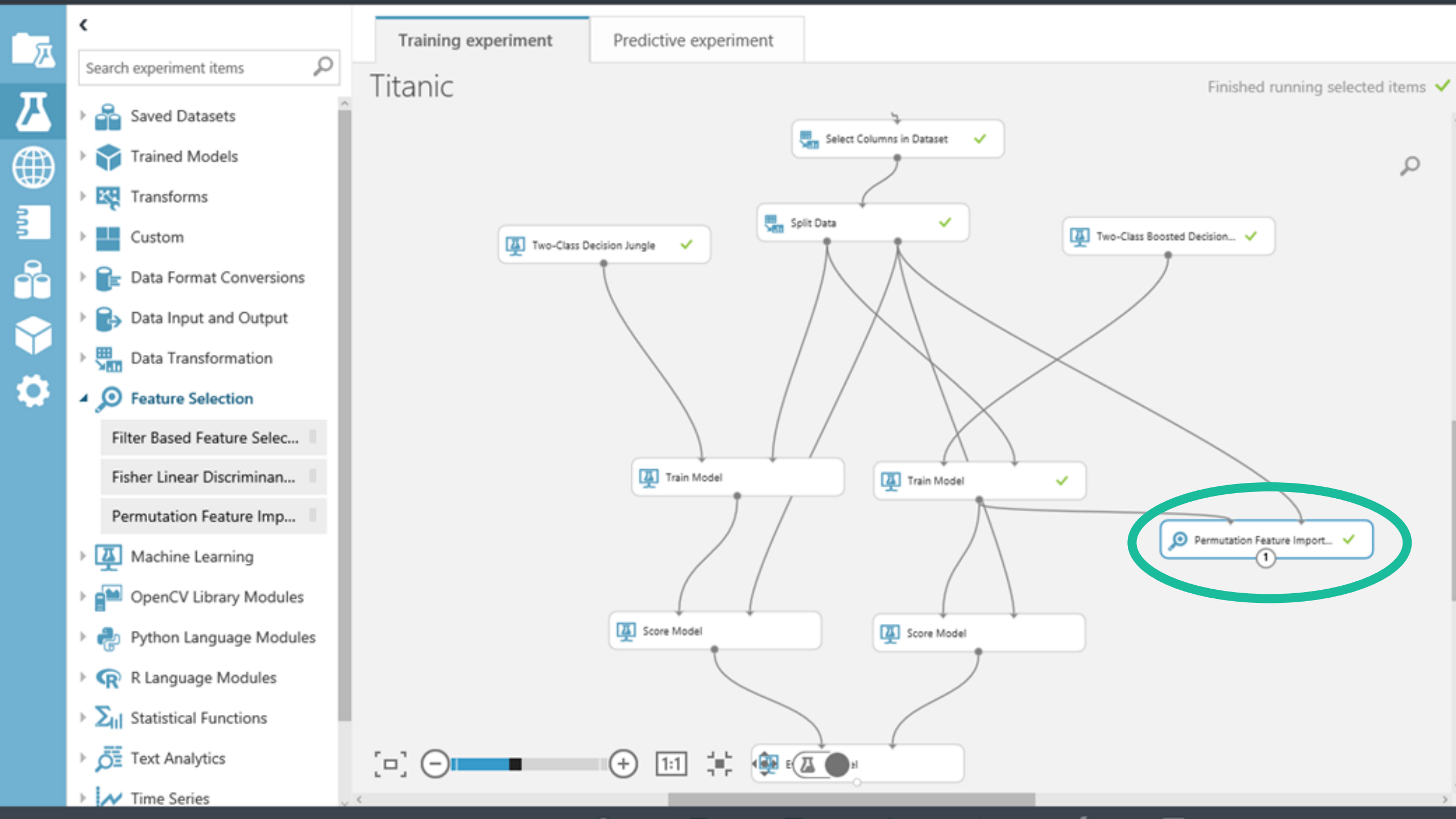
Wybór odpowiednich kolumn

Wybór odpowiednich kolumn

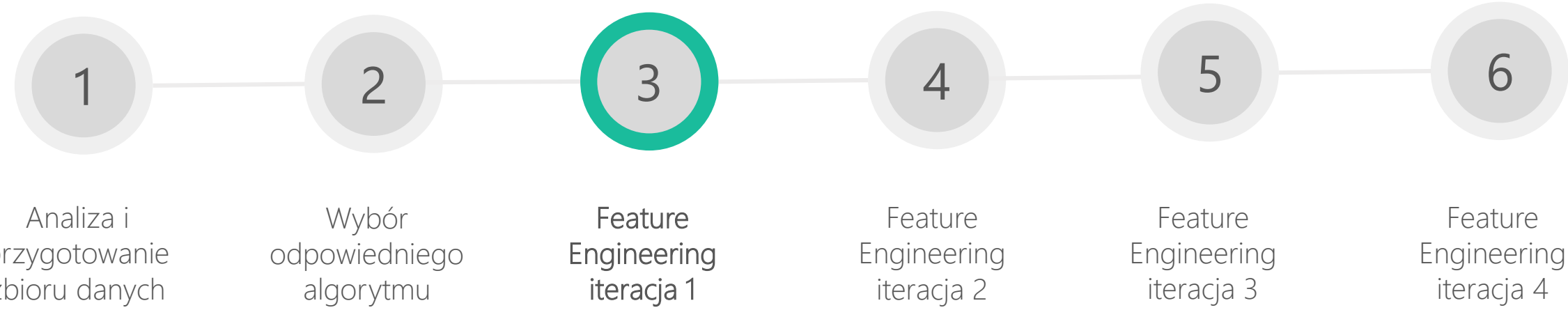
Permutation Feature Importance – pozwala wybrać atrybuty mające największy wpływ na poprawną klasyfikację zmiennej modelowanej.

Po wstępnej analizie na placu boju zostają:

- Survived
- Sex
- Age
- Pclass
- Fare


Search experiment items
Saved Datasets
Trained Models
Transforms
Custom
Data Format Conversions
Data Input and Output
Data Transformation
Feature Selection
Filter Based Feature Selec...
Fisher Linear Discriminan...
Permutation Feature Imp...
Machine Learning
OpenCV Library Modules
Python Language Modules
R Language Modules
Statistical Functions
Text Analytics
Time Series
Training experiment
Predictive experiment
Titanic
Finished running selected items
Select Columns in Dataset
Split Data
Two-Class Decision Jungle
Two-Class Boosted Decision...
Train Model
Train Model
Permutation Feature Import...
1
Score Model
Score Model
1:1

Proces budowy modelu predykcyjnego



Na których zmiennych się skupić?

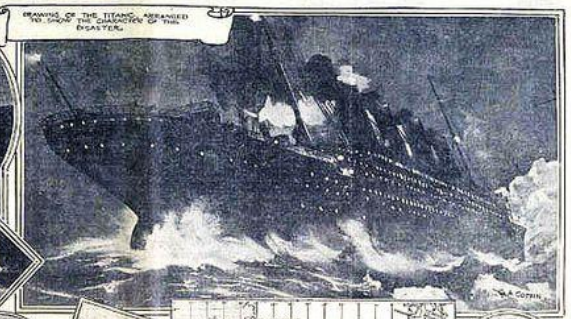
DIRECTORY FOR ADVERTISERS WITH DESCRIPTIVE INDEX OF PAGES 11, 12 and 13.

THE NEW YORK HERALD.

THE WEATHER. Clearing, slightly warmer.

WHOLE NO. 27,630. NEW YORK, TUESDAY, APRIL 16, 1912—TWENTY-EIGHT PAGES. PRICE THREE CENTS.

THE TITANIC SINKS WITH 1,800 ON BOARD; ONLY 675, MOSTLY WOMEN AND CHILDREN, SAVED



MAP SHOWING LOCATION OF THE TITANIC AT THE TIME OF THE ACCIDENT

MOST APPALLING DISASTER IN MARINE HISTORY OCCURS WHEN WORLD'S LARGEST STEAMSHIP STRIKES GIGANTIC ICEBERG AT NIGHT

Scores of World's Most Widely Known Persons, Including Colonel John Jacob Astor and His Wife, William T. Stead and Isidor Straus Among Those Whose Fate Is in Doubt.

The Carpathia's Survivors List

A partial list of the survivors now on board the Carpathia, bound for this port, will be found on page 2 of this issue.

J. Bruce Ismay, Head of the White Star Line; Mr. and Mrs. George D. Widener, the Countess of Rothes and Mrs. Henry B. Harris Are Some of Those Saved.

In the darkness of night and in water two miles deep the Titanic, newest of the White Star fleet and greatest of all ocean steamships, sank to the bottom of the sea at twenty minutes past two o'clock yesterday morning.

Despatches received late last night from the Cape Race Wireless station in Newfoundland and addressed reluctantly inside at the same time by the New York officials of the White Star Company warrant the fear that of the 2,200 persons who were aboard the great vessel when she struck her mortal wound in collision with an iceberg more than 1,500 have gone to their death in her shattered hulk, while 675, most of whom are women and children, have been saved.

Should these grim figures be verified, the loss of the Titanic—smallest, most powerful, greatest of all the ocean fleet—while speeding westward on her maiden voyage will take rank in maritime history as the most terrible of all recorded disasters of the sea.

A despatch from St. Johns, N. F., received at fifteen minutes past two o'clock [illegible] morning given rise to the hope that the steamship Virginian may have some of the Titanic's survivors on board. The message said she would bring to St. Johns such survivors as she "may rescue." Hope arises from the fact that the steamer is putting in here at all, which she would scarcely need to do were there not some humane necessity for that course. She was outward bound for Liverpool.

One point is known from which may be derived a sad satisfaction. In a desperate situation where the risk [illegible] region of all was not possible the women and children were cared for first. These were sent away in the first of the [illegible] boats launched from the sinking ship, the only boats apparently which did not share the fate of the steamship vessel.

America and Britain are spared the horrors that attended the sinking of the French ship La Bourgogne, in [illegible] 1898, when the women and children were trampled under foot and cut down with knives.

CAPTAIN E. J. SMITH

CLARENCE MOORE

WILLIAM T. STEAD

MRS. JAMES WARNER MARVIN

It was learned late last night that the full text of the message New York aboard the Carpathia, of the Cunard line, the loss of life from the steamship Olympic reporting the sinking of the Titanic, only indicated would be 1,400 persons.

The full text of the despatch received from the Olympic is as follows:—

"Carpathia reached Titanic's position at daybreak. Titanic sent About seventy minutes past two o'clock in the morning, in 41.16 north latitude, 50.14 west longitude. All her boats accounted for, containing about 675 souls saved, crew and passengers included. Nearly all based women and children. Leyland Line Californian remained and searching exact position of disaster. Less likely 1,600 souls."

By midnight Bowling Green in front of the White Star offices

Feature Engineering – iteracja numer 1

1. Uzupełnienie brakujących wartości w kolumnie „Age” – średnia wprowadza „szum”.
2. Dodanie nowych kolumn opartych o atrybuty mające największy wpływ na zmienną modelowaną.

Properties Project

R Script

```
19
20 #Osoby z tytułem Miss i Mister, to w większości dzieci.
21 #Master-średia wieku 4.5.
22 #Miss to częściowo młode kobiety, a częściowo starsze panie. Rozróżnia się je dzięki Parch. Jeżeli liczba rodziców = 0, to m
23 setDT(plik)[grepl("Master", Name) & is.na(Age), Age := 4.5]
24 setDT(plik)[grepl("Miss", Name) & is.na(Age) & Parch == 0, Age := 28]
25
26 #Sir, Mr, Ms i Mrs to dojrzałe osoby
27 setDT(plik)[grepl("Mr\\.", Name) & is.na(Age), Age := 31]
28 setDT(plik)[grepl("Sir\\.", Name) & is.na(Age), Age := 49]
29 setDT(plik)[grepl("Ms\\.", Name) & is.na(Age), Age := 28]
30 setDT(plik)[grepl("Mrs\\.", Name) & is.na(Age), Age := 36]
31
32 #Dr to dojrzały facet. Średnia wieku to 42 lata.
33 setDT(plik)[grepl("Dr\\.", Name) & is.na(Age), Age := 42]
34
35 #Reszta kobiet z tytułem Miss to starsze kobiety, którym nie udało się przeżyć.
36 setDT(plik)[grepl("Miss", Name) & is.na(Age), Age := 40]
37
38 #Reszta osób z niezdefiniowanym wiekiem
39 plik$Age[is.na(plik$Age) & plik$Sex == "male"] <- srednia.wieku.mezczyzny
40 plik$Age[is.na(plik$Age) & plik$Sex == "female"] <- srednia.wieku.kobiety
41
42 #Tworzę przedziały wiekowe. W dalszej kolejności będę sprawdzał ich użyteczność
```

Dodanie nowych kolumn

1. „Family.Size" – tworzę ją przez dodanie do siebie dwóch istniejących wartości: „Parch" i „SibSp", oraz liczby 1, odpowiadającej za daną osobę. Intuicyjnie uznaję, że wielkość rodziny mogła mieć znaczący wpływ na to czy komuś udało się przeżyć, czy też nie.

2. „Age.Range" – jest to zmienna kategoryczna, która przypisuje pasażera do jednej z czterech kategorii wiekowych: „Bobas", „Dzieciak", „Nastolatek", „Dorosly".

```
plik$Age.Range <- cut(plik$Age,c(0,6,12,18,Inf),labels = c( „Bobas","Dzieciak","Nastolatek","Dorosly"))
```

3. „MPC" – zmienna, którą chcę „uwypuklić" szanse na przeżycie dzieci i osób z pierwszej klasy. Jest ona wynikiem mnożenia wieku danej osoby i klasy, w której podróżowała, np.: 5-letnie dziecko podróżujące w pierwszej klasie (wynik = 5), miało dużo większe szanse na przetrwanie katastrofy niż 70 letni pan podróżujący klasą trzecią (wynik = 210).

Wyniki i wnioski
iteracja numer 1

Experiments - Microsoft A... | Experiments - Microsoft A...

https://studio.azureml.net/Home/ViewWorkspaceCached/28c7c5dcee5a4732b758c90727385c03?#Workspaces/Experiments/Experiment/28c7c5dcee5a4732b

110%

Microsoft Azure Machine Learning Studio

Mateusz Grzyb-Free-Work...

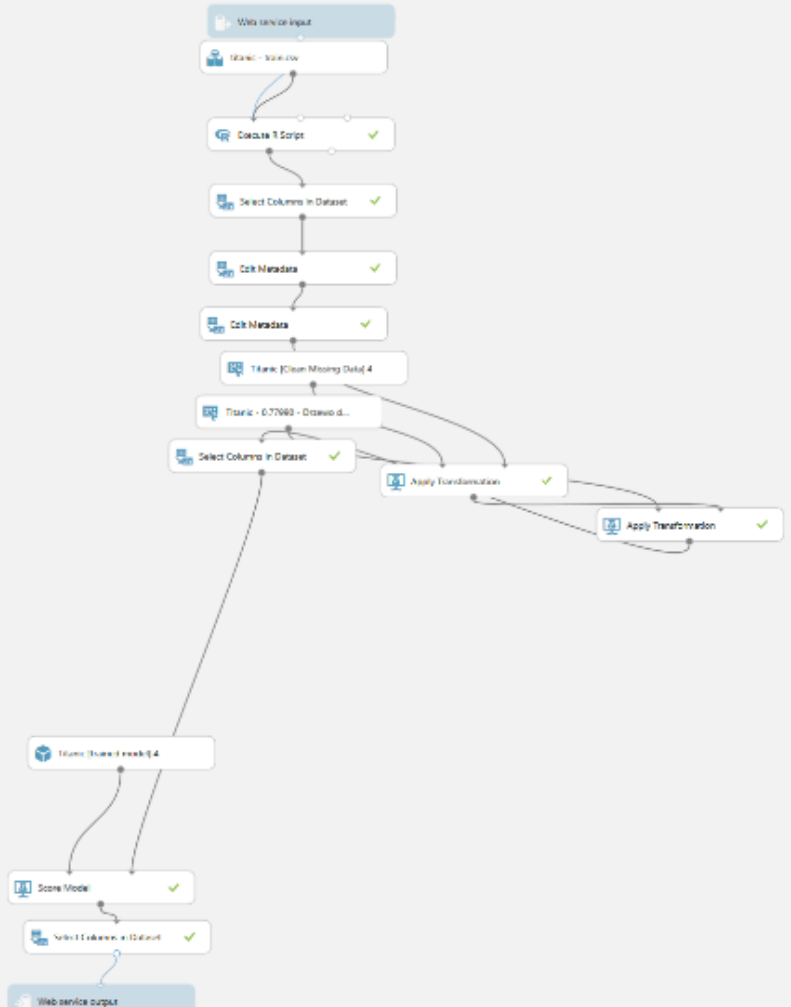
Search experiment items

- Saved Datasets
- Trained Models
- Transforms
- Data Format Conversions
- Data Input and Output
- Data Transformation
  - Filter
  - Learning with Counts
  - Manipulation
    - Add Columns
    - Add Rows
    - Apply SQL Transform...
    - Clean Missing Data
    - Convert to Indicator ...
    - Edit Metadata

Training experiment | Predictive experiment

Titanic [Predictive Exp.]

Finished running



Properties | Project

Experiment Properties

START TIME	2/5/2017...
END TIME	2/5/2017...
STATUS CODE	Finished
STATUS DETAILS	None

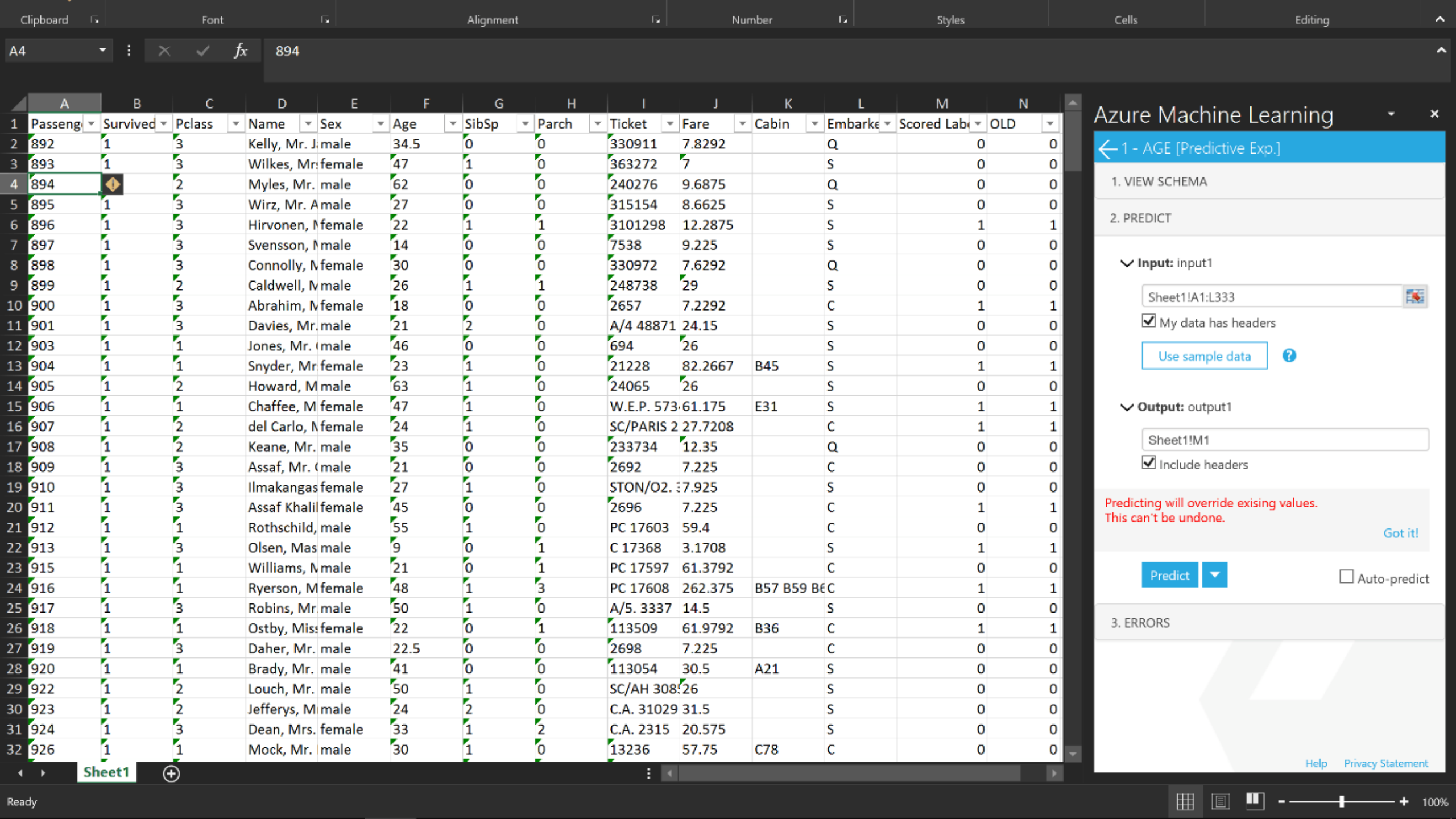
Summary

Enter a few sentences describing your experiment (up to 140 characters).

Description

Enter the detailed description for your experiment.

Quick Help



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Passeng	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarke	Scored Lab	OLD
2	892	1	3	Kelly, Mr. J	male	34.5	0	0	330911	7.8292		Q	0	0
3	893	1	3	Wilkes, Mr	female	47	1	0	363272	7		S	0	0
4	894		2	Myles, Mr.	male	62	0	0	240276	9.6875		Q	0	0
5	895	1	3	Wirz, Mr. A	male	27	0	0	315154	8.6625		S	0	0
6	896	1	3	Hirvonen, M	female	22	1	1	3101298	12.2875		S	1	1
7	897	1	3	Svensson, M	male	14	0	0	7538	9.225		S	0	0
8	898	1	3	Connolly, M	female	30	0	0	330972	7.6292		Q	0	0
9	899	1	2	Caldwell, M	male	26	1	1	248738	29		S	0	0
10	900	1	3	Abrahim, M	female	18	0	0	2657	7.2292		C	1	1
11	901	1	3	Davies, Mr.	male	21	2	0	A/4 48871	24.15		S	0	0
12	903	1	1	Jones, Mr. (	male	46	0	0	694	26		S	0	0
13	904	1	1	Snyder, Mr	female	23	1	0	21228	82.2667	B45	S	1	1
14	905	1	2	Howard, M	male	63	1	0	24065	26		S	0	0
15	906	1	1	Chaffee, M	female	47	1	0	W.E.P. 573	61.175	E31	S	1	1
16	907	1	2	del Carlo, M	female	24	1	0	SC/PARIS 2	27.7208		C	1	1
17	908	1	2	Keane, Mr.	male	35	0	0	233734	12.35		Q	0	0
18	909	1	3	Assaf, Mr. (	male	21	0	0	2692	7.225		C	0	0
19	910	1	3	Ilmakangas	female	27	1	0	STON/O2. 3	7.925		S	0	0
20	911	1	3	Assaf Khalil	female	45	0	0	2696	7.225		C	1	1
21	912	1	1	Rothschild,	male	55	1	0	PC 17603	59.4		C	0	0
22	913	1	3	Olsen, Mas	male	9	0	1	C 17368	3.1708		S	1	1
23	915	1	1	Williams, M	male	21	0	1	PC 17597	61.3792		C	0	0
24	916	1	1	Ryerson, M	female	48	1	3	PC 17608	262.375	B57 B59 B6	C	1	1
25	917	1	3	Robins, Mr	male	50	1	0	A/5. 3337	14.5		S	0	0
26	918	1	1	Ostby, Miss	female	22	0	1	113509	61.9792	B36	C	1	1
27	919	1	3	Daher, Mr.	male	22.5	0	0	2698	7.225		C	0	0
28	920	1	1	Brady, Mr.	male	41	0	0	113054	30.5	A21	S	0	0
29	922	1	2	Louch, Mr.	male	50	1	0	SC/AH 308	26		S	0	0
30	923	1	2	Jefferys, M	male	24	2	0	C.A. 31029	31.5		S	0	0
31	924	1	3	Dean, Mrs.	female	33	1	2	C.A. 2315	20.575		S	0	0
32	926	1	1	Mock, Mr.	male	30	1	0	13236	57.75	C78	C	0	0

kaggle Host Competitions Datasets Kernels Jobs Community MateuszGrzyb Logout

Knowledge • 6,155 teams

Titanic: Machine Learning from Disaster

Fri 28 Sep 2012 — Tue 7 Jan 2020 (35 months to go)

Dashboard

Public Leaderboard - Titanic: Machine Learning from Disaster

This leaderboard is calculated or
The final results will be based or

#	Δ1w	Team Nam	sion UTC (Best – Last Submission)
1	—	Gang of	v 2016 16:14:57
2	—	Prabhud	2016 17:27:41
3	—	MikhailO	2016 15:09:18
4	—	腰力	2016 02:51:29

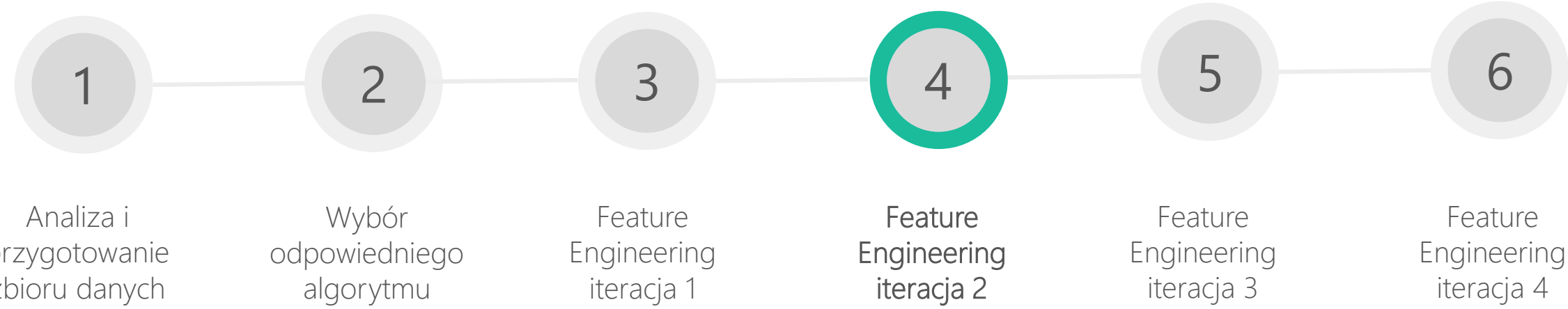
Making a call to Watson to check this for you...

Wyniki i wnioski - iteracja numer 1

Pierwsze wnioski wyglądały następująco:

- Zmienna „MPC” nie wpłynęła pozytywnie na wynik.
- Zmienne „Family.Size”, oraz „Age.Range” wpłynęły pozytywnie na wynik osiągany na zbiorze testowym.
- Spośród testowanych algorytmów uczenia maszynowego najlepsze wyniki osiągał „Two-Class Decision Jungle”.

Proces budowy modelu predykcyjnego



Feature Engineering – iteracja numer 2

1. Co o sytuacji danej osoby mówi jej tytuł?
   - „Master" to młody kawaler.
   - „Mml" nosiły niezamężne Francuzki i jest to odpowiednik dla angielskiego „Miss".
2. Jeszcze lepsza estymacja wieku danej osoby.

Properties Project

R Script

```r

plik$Title <- as.factor("")
setDT(plik)[grepl("Master\\.", Name), Title := "Master"]
setDT(plik)[grepl("Miss\\.", Name), Title := "Miss"]
setDT(plik)[grepl("Mr\\.", Name), Title := "Mr"]
setDT(plik)[grepl("Mrs\\.", Name), Title := "Mrs"]
setDT(plik)[grepl("Sir\\.", Name), Title := "Sir"]
setDT(plik)[grepl("Ms\\.", Name), Title := "Ms"]
setDT(plik)[grepl("Don\\.", Name), Title := "Don"]
setDT(plik)[grepl("Dona\\.", Name), Title := "Dona"]
setDT(plik)[grepl("Rev\\.", Name), Title := "Rev"]
setDT(plik)[grepl("Dr\\.", Name), Title := "Dr"]
setDT(plik)[grepl("Rev\\.", Name), Title := "Rev"]
setDT(plik)[grepl("Major\\.", Name), Title := "Major"]
setDT(plik)[grepl("Lady\\.", Name), Title := "Lady"]
setDT(plik)[grepl("Mme\\.", Name), Title := "Mme"]
setDT(plik)[grepl("Mlle\\.", Name), Title := "Mlle"]
setDT(plik)[grepl("Col\\.", Name), Title := "Col"]
setDT(plik)[grepl("Capt\\.", Name), Title := "Capt"]
setDT(plik)[grepl("Countess\\.", Name), Title := "Countess"]
setDT(plik)[grepl("Jonkheer\\.", Name), Title := "Jonkheer"]

#Osoby z tytułem Miss i Mister, to w większości dzieci.
#Master-średnia wieku 4.5.
```

Jeszcze lepsza estymacja wieku danej osoby

1. Wyestymować wartość wieku z pomocą algorytmu regresyjnego, bezpośrednio w Azure ML.

2. Wyestymować wartość wieku z pomocą algorytmu regresyjnego z poziomu kodu R, lub Python.

3. Użyć podejścia niekonwencjonalnego i podzielić jeden duży eksperyment na dwa mniejsze: jeden dla pasażerów, którzy mają podany wiek, oraz drugi dla wszystkich których wieku nie znamy.

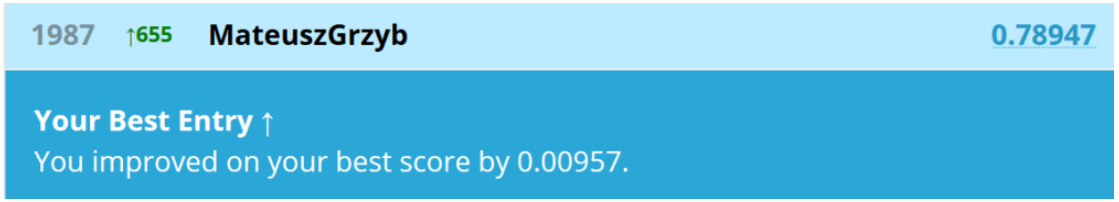
Jeszcze lepsza estymacja wieku danej osoby

1. Wyestymować wartość wieku z pomocą algorytmu regresyjnego, bezpośrednio w Azure ML.
2. Wyestymować wartość wieku z pomocą algorytmu regresyjnego z poziomu kodu R, lub Python.
3. Użyć podejścia niekonwencjonalnego i podzielić jeden duży eksperyment na dwa mniejsze: jeden dla pasażerów, którzy mają podany wiek, oraz drugi dla wszystkich których wieku nie znamy.

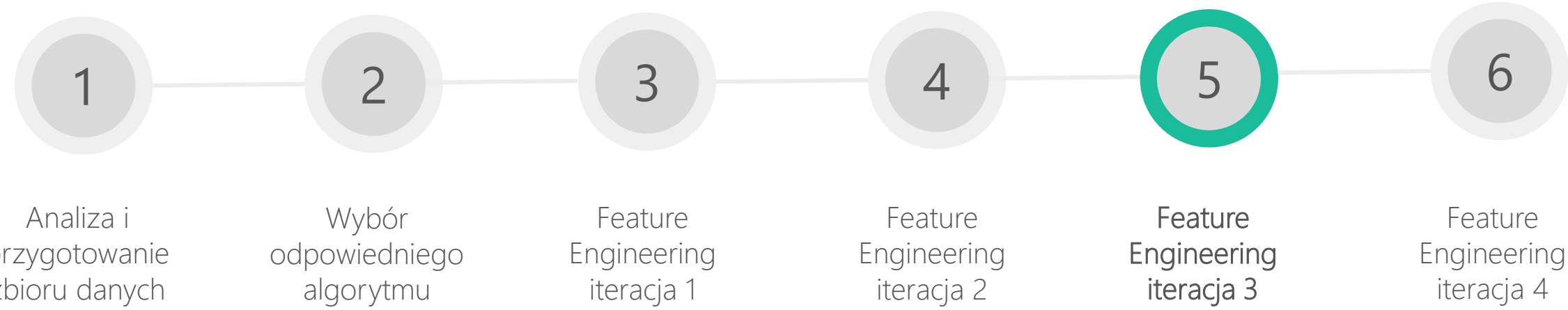
Wyniki i wnioski - iteracja numer 2

Wynik znacząco podskoczył. Już jest lepszy od najlepszego uzyskanego przez autorów tutoriali dostępnych na Kaggle.

- Kolumna „Title" znacząco wpływa na wynik.
- Kolumna „Age" jest kluczem do sukcesu w konkursie.
- Dodatkowy model regresyjny poprawia wynik na zbiorze uczącym.



Proces budowy modelu predykcyjnego



Feature Engineering – iteracja numer 3

1. Uzupełnienie brakujących wartości w kolumnie „Fare”.
2. Kolejny pomysł na poprawienie dokładności estymacji wieku danego pasażera: użycie tytułu jaki posiadał.
3. Im większa rodzina tym mniejsze szanse na przeżycie?

Uzupełnienie brakujących wartości w kolumnie „Fare”.

Kilku pasażerów nie ma podanej wartości ceny biletu.

Są to głównie mężczyźni płynący trzecią klasą, wyruszający z portu „Southampton”.

Z pomocą R, uzupełniam te braki medianą ceny biletu dla danej grupy (np. mediana opłaty za bilet, pasażera trzeciej klasy płynącego z Southampton, podróżującego samotnie).

Użycie tytułu jaki posiadała dana osoba do estymacji wieku

Po ponownym zastosowaniu eksploracyjnej analizy danych dostrzegłem ciekawą prawidłowość: spośród wszystkich osób, żaden inny tytuł nie wskazuje wieku pasażera z taką dokładnością jak „Master" i „Dr". Pierwszy z nich jest przypisany w zdecydowanej większości dla dzieci płci męskiej, drugi natomiast dla starszych pasażerów.

Zdecydowałem się przypisać wartości dla wszystkich osób posiadających wymienione tytuły, a jednocześnie nie posiadających podanego wieku, „na sztywno". Posiłkuję się w tym przypadku wartościami średnimi.

Im większa rodzina tym mniejsze szanse na przeżycie?

1. Jako założenie przyjąłem, że mała rodzina, to taka która posiada mniej niż 3 osoby.
2. Dodałem zatem zmienną „Family.ID", która jednoznacznie pozwala mi identyfikować osoby przynależące do danej rodziny. Utworzyłem ją z nazwiska danej osoby, oraz wartości „Family.Size".
3. By uniknąć zbyt dużej liczby zmiennych kategorycznych postanowiłem zastąpić wszystkie wartości „Family.ID" dla rodzin mniejszych niż 3, wartością „Small". Pozwoliło mi to zachować rozsądną liczbę 34 kategorii dla tej kolumny.
4. By podkreślić wielkość rodziny i jej wpływ na wartość „Survived", dodałem kolumnę „Family.Size.P". Przyjmowała ona jedną z dwóch wartości: „Small", lub „Big".

Microsoft Azure Machine Learning Studio

Mateusz Grzyb-Free-Work...

Training Properties Project

Search

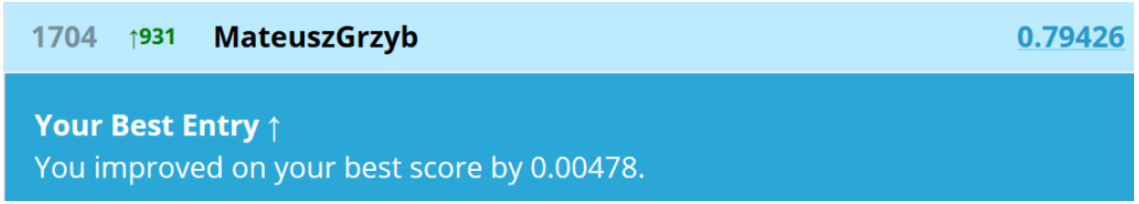
R Script

```r
#Master-średia wieku 4.5.
setDT(plik)[grepl("Master", Name) & is.na(Age), Age := 3.5]

#Dr to dojrzały facet. Średnia wieku to 42 lata.
setDT(plik)[grepl("Dr\\.", Name) & is.na(Age), Age := 42]

#Pasażerowie bez podanej ceny biletu będą mieć estymowaną cenę.
plik$Fare[is.na(plik$Fare)] <- median(plik$Fare[plik$Pclass == "3" & plik$Embarked == "S" & plik$Sex == "male" & plik$SibSp

#Wyłuskanie nazwiska
plik$Surname <- sapply(as.character(plik$Name), FUN = function(x) {strsplit(x, split='[,.]')[[1]][1]})
plik$Surname <- as.factor(plik$Surname)

#Dodaję ID rodziny. Jest teoria, która mówi o tym że rodziny wieloosobowe miały mniejsze szanse by przerzyć.
plik$Family.ID <- paste(plik$Family.Size, plik$Surname, sep = "")
famIDs <- data.frame(table(plik$Family.ID))
famIDs <- famIDs[famIDs$Freq <= 2,]
plik$Family.ID[plik$Family.ID %in% famIDs$Var1] <- 'Small'
plik$Family.ID <- factor(plik$Family.ID)
plik$Family.Size.P <- 'Big'
plik$Family.Size.P[plik$Family.ID=='Small'] <- 'Small'
plik$Family.Size.P <- as.factor(plik$Family.Size.P)

```

Wyniki i wnioski - iteracja numer 3

Wnioski po trzeciej iteracji wyglądają następująco:

- Nowa kolumna – „Family.ID" – pozytywnie wpłynęła na wynik.
- Zmniejszenie liczby zmiennych w kolumnie „Family.ID" podniosło efektywność modelu.
- Równocześnie z pozytywnym wpływem „Family.ID", zaobserwowałem odwrotny efekt w przypadku zmiennej „Family.Size.P". Okazała się ona zupełnie niepotrzebna i jedynie zaniżała osiągany wynik.



Proces budowy modelu predykcyjnego

1. Analiza i przygotowanie zbioru danych
2. Wybór odpowiedniego algorytmu
3. Feature Engineering iteracja 1
4. Feature Engineering iteracja 2
5. Feature Engineering iteracja 3
6. **Feature Engineering iteracja 4**

Feature Engineering – iteracja numer 4

1. Powrót do pomysłów z iteracji numer 2.
2. Zróżnicowanie liczebności zbiorów.

Powrót do pomysłów z iteracji numer 2

Jeszcze lepsza estymacja wieku danej osoby poprzez użycie podejścia niekonwencjonalnego i podzielenie jednego dużego eksperymentu na dwa mniejsze:

1. Jeden dla pasażerów, którzy mają podany wiek (zawiera wszystkich zmienne predykcyjne, które sprawdzały się do tej pory, ze zmienną „Age" włącznie).
2. Drugi dla wszystkich których wieku nie znamy (wszystkie zmienne z wyjątkiem zmiennej „Age" oraz „Age.Range").

Zróżnicowanie liczebności zbiorów

W poprzednim kroku utworzyłem 2 nowe eksperymenty Azure ML, oparte na dwóch różnych zbiorach.

Pierwotnie miałem do dyspozycji 2 zbiory: uczący (891 obserwacji; 719 ma podany wiek; 172 nie ma podanego wieku) i testowy (418 obserwacji).

Zróżnicowanie liczebności zbiorów

Zamiast dzielić zbiór uczący na podzbiory o liczebności: 719 i 172, dokonuję podziału:

1. Zbiór uczący eksperymentu A: 719 (podany wiek) – użyłem do wszystkich przypadków gdzie był podany wiek pasażera.
2. Zbiór uczący eksperymentu B: 891 (bez podanego wieku) – użyłem do wszystkich przypadków gdzie był podany wiek pasażera..

Microsoft Azure Machine Learning Studio

Mateusz Grzyb-Free-Work...

Training experiment | Predictive experiment

Properties | Project

Titanic - bez podanego wieku › Evaluate Model › Evaluation results

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
71	31	0.858	0.910	0.5	0.884

False Positive	True Negative	Recall	F1 Score
7	158	0.696	0.789

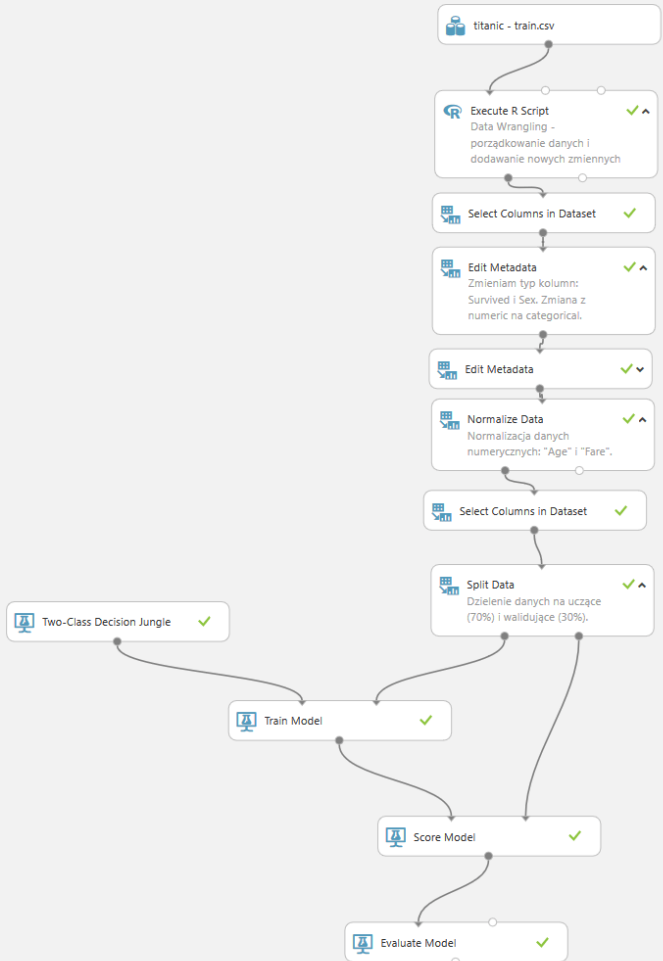
Positive Label	Negative Label
1	0

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	35	1	0.135	0.745	0.507	0.972	0.343	0.710	0.994	0.001
(0.800,0.900]	15	0	0.191	0.801	0.654	0.980	0.490	0.759	0.994	0.001
(0.700,0.800]	3	1	0.206	0.809	0.675	0.964	0.520	0.769	0.988	0.005
(0.600,0.700]	5	0	0.225	0.828	0.716	0.967	0.569	0.787	0.988	0.005
(0.500,0.600]	13	5	0.292	0.858	0.789	0.910	0.696	0.836	0.958	0.024
(0.400,0.500]	10	16	0.390	0.835	0.786	0.779	0.794	0.871	0.861	0.098
(0.300,0.400]	6	8	0.442	0.828	0.791	0.737	0.853	0.899	0.812	0.137
(0.200,0.300]	0	10	0.479	0.790	0.757	0.680	0.853	0.892	0.752	0.189
(0.100,0.200]	8	62	0.742	0.588	0.633	0.480	0.931	0.899	0.376	0.522
(0.000,0.100]	7	62	1.000	0.382	0.553	0.382	1.000	1.000	0.000	0.884

1 - AGE

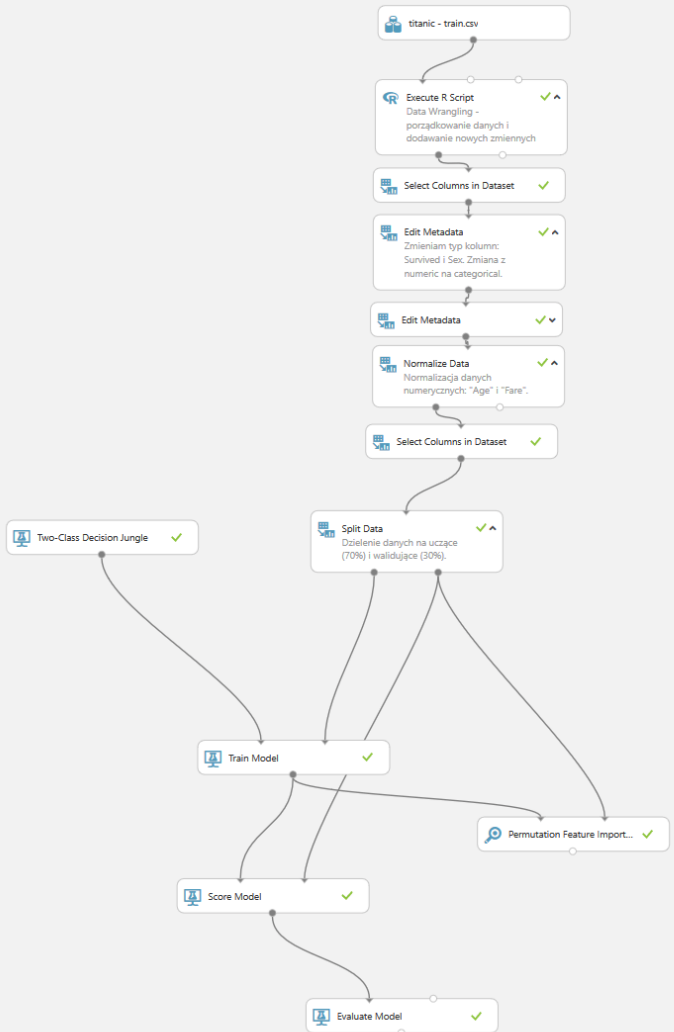
Finished running

Draft saved at 17:29:24



1 - NA

Finished running



Wyniki i wnioski - iteracja numer 4

Końcowe Wnioski:

- Zmiana podejścia przyniosła oczekiwane rezultaty.
- Mój model regresyjny jednak nie był tak dobry jak mi się pierwotnie wydawało i wnosił sporo „szumu" do danych. Nauczka na przyszłość dla mnie, by równolegle testować przynajmniej dwa alternatywne podejścia do budowania modelu.
- Im dalej w las, tym więcej drzew. Każda kolejna próba poprawy modelu była większym wyzwaniem niż poprzednia.
- Często najlepsze rezultaty dają najprostsze metody.

Podsumowanie

Podsumowanie

- Azure Machine Learning Studio – szybkość, prostota działania.
- Jeśli oczekujesz wysokiej skalowalności, szybkości i customizacji: HDInsight, R Server, DSVM + R.
- Czasem Decision Jungle daje lepsze rezultaty niż Boosted Decision Tree.
- Finalnie osiągnięty przeze mnie wynik dał miejsce w okolicy 500 spośród 6000 zespołów, a więc jest to pierwsze 10% spośród wszystkich notowanych zespołów.

Pytania do Was

Który algorytm dał najlepsze rezultaty?

Odp: Decision Jungle

Jakim parametrem mierzyłem dokładność algorytmów?

Odp: Dokładność (ang. *Accuracy*)

Jakiego typu był problem z którym się mierzyłem?

Odp: Klasyfikacji

Pytania do mnie

Materiały: mateuszgrzyb.pl/MAUG

Kontakt: m.grzyb@outlook.com

Machine Learning in ML Studio

Anomaly Detection
- One-class Support Vector Machine
- Principal Component Analysis-based Anomaly Detection
- Time Series Anomaly Detection*

Classification

Two-class Classification
- Averaged Perceptron
- Bayes Point Machine
- Boosted Decision Tree
- Decision Forest
- Decision Jungle
- Logistic Regression
- Neural Network
- Support Vector Machine

Multi-class Classification
- Decision Forest
- Decision Jungle
- Logistic Regression
- Neural Network
- One-vs-all

Clustering
- K-means Clustering

Recommendation
- Matchbox Recommender

Regression
- Bayesian Linear Regression
- Boosted Decision Tree
- Decision Forest
- Fast Forest Quantile Regression
- Linear Regression
- Neural Network Regression
- Ordinal Regression
- Poisson Regression

Statistical Functions
- Descriptive Statistics
- Hypothesis Testing T-Test
- Linear Correlation
- Probability Function Evaluation

Text Analytics
- Feature Hashing
- Named Entity Recognition
- Vowpal Wabbit

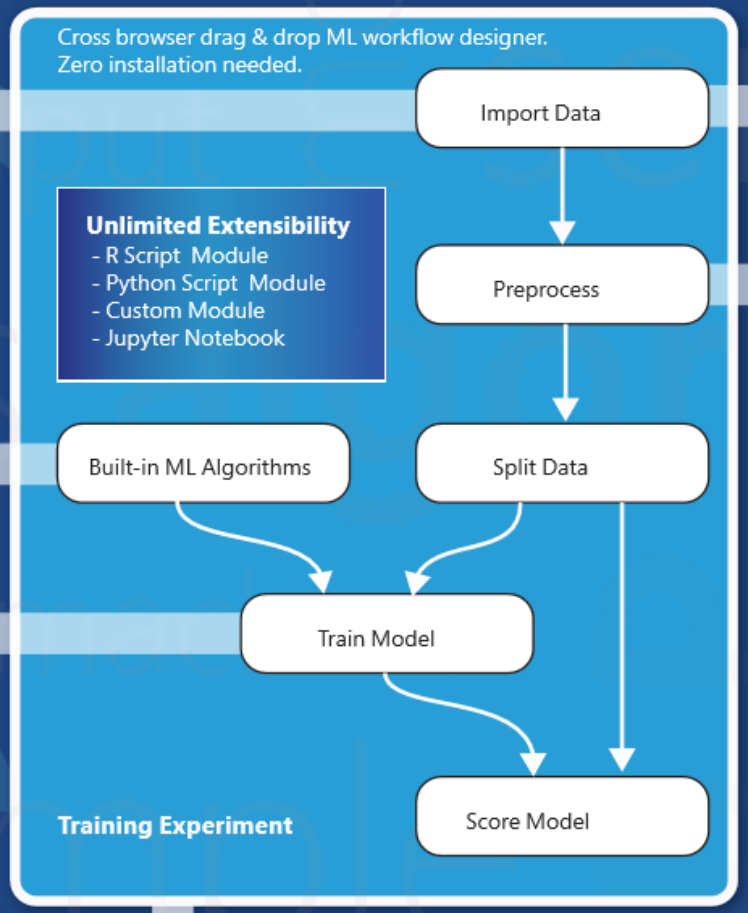
Computer Vision
- OpenCV Library

https://studio.azureml.net

Guest Access Workspace: Free trial access without logging in.
Free Workspace: Free persisted access, no Azure subscription needed.
Standard Workspace: Full access with SLA under an Azure subscription.

Data/Model Visualization
- Scatterplots
- Bar Charts
- Box plots
- Histogram
- R and Python Plotting Libraries
- REPL with Jupyter Notebook
- ROC, Precision/Recall, Lift
- Confusion Matrix
- Decision Tree*

Training
- Cross Validation
- Retraining
- Parameter Sweep



Unlimited Extensibility
- R Script Module
- Python Script Module
- Custom Module
- Jupyter Notebook

Data Source
- Azure Blob Storage
- Azure SQL DB
- Azure SQL DW*
- Azure Table
- Desktop Direct Upload
- Hadoop Hive Query
- Manual Data Entry
- OData Feed
- On-prem SQL Server*
- Web URL (HTTP)

Data Format
- ARFF
- CSV
- SVMLight
- TSV
- Excel
- ZIP

Data Preparation
- Clean Missing Data
- Clip Outliers
- Edit Metadata
- Feature Selection
- Filter
- Learning with Counts
- Normalize Data
- Partition and Sample
- Principal Component Analysis
- Quantize Data
- SQLite Transformation
- Synthetic Minority Oversampling Technique

Enterprise Grade Cloud Service
- SLA: 99.95% Guaranteed Up-time
- Azure AD Authentication
- Compute at Large Scale
- Multi-geo Availability
- Regulatory Compliance*

One-click Operationalization

Predictive Experiment

Make Prediction with Elastic APIs
- Request-Response Service (RRS)
- Batch Execution Service (BES)
- Retraining API

Community
- Gallery (http://gallery.azureml.net)
- Samples & Templates
- Workspace Sharing and Collaboration
- Live Chat & MSDN Forum Support

Search experiment items

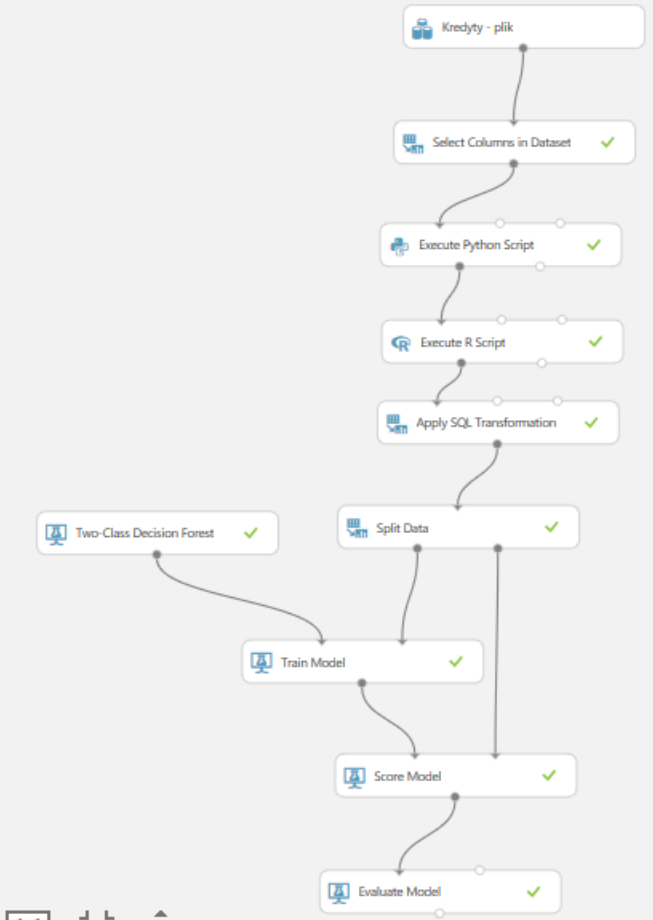
- Saved Datasets
- Trained Models
- Transforms
- Custom
- Data Format Conversions
- Data Input and Output
- Data Transformation
- Feature Selection
- Machine Learning
- OpenCV Library Modules
- Python Language Modules
- R Language Modules
- Statistical Functions
- Text Analytics
- Time Series
- Web Service
- Deprecated

Training experiment | Predictive experiment

Klasyfikacja kredytobiorców

Finished running

Draft saved at 10:53:47



Mini Map

Properties | Project

Experiment Properties

START TIME	11/18/20...
END TIME	11/18/20...
STATUS CODE	Finished
STATUS DETAILS	None

Prior Run

Summary

Enter a few sentences describing your experiment (up to 140 characters).

Description

Enter the detailed description for your experiment.

Quick Help

NEW | RUN HISTORY | SAVE | SAVE AS | DISCARD CHANGES | RUN | SET UP WEB SERVICE | PUBLISH TO GALLERY