Fifty shades of....

Automated Machine Learning

Rui Quintino, Data Research @ DevScope
Dewan Fayzur, Data Scientist @ DevScope

Rui Quintino

Data R&D @ DevScope

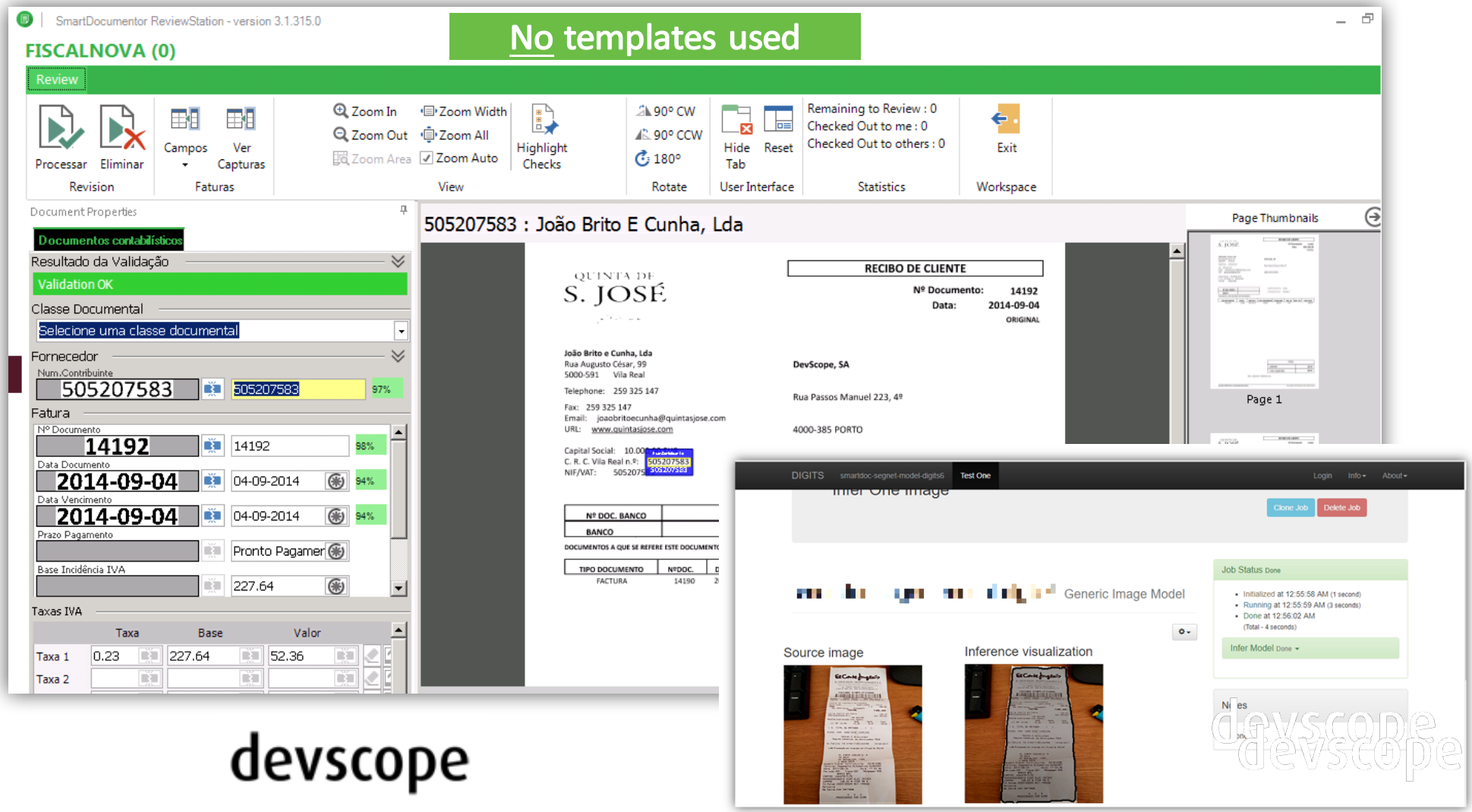
#PowerBI #SQLServer #Web
#Analytics #Azure #Microsoft

#MachineLearning #AutoML
#R #Linux #Dataiku #Docker
#Python #Coaching #Learning

also very often a #DataSkeptic ☺

twitter.com/rquintino
rquintino.wordpress.com
rquintino@gmail.com

"jack of all trades (and master of none)"
1. a person who can do many different types of work but who is not (necessarily…) very competent at any of them…


SmartDocumentor ReviewStation - version 3.1.315.0
No templates used
FISCALNOVA (0)
Review
Processar
Eliminar
Revision
Campos
Ver Capturas
Faturas
Zoom In
Zoom Out
Zoom Area
Zoom Width
Zoom All
Zoom Auto
Highlight Checks
View
90° CW
90° CCW
180°
Rotate
Hide Tab
Reset
User Interface
Remaining to Review : 0
Checked Out to me : 0
Checked Out to others : 0
Statistics
Exit
Workspace
Document Properties
Documentos contabilísticos
Resultado da Validação
Validation OK
Classe Documental
Selecione uma classe documental
Fornecedor
Num.Contribuinte
505207583
505207583
97%
Fatura
Nº Documento
14192
14192
98%
Data Documento
2014-09-04
04-09-2014
94%
Data Vencimento
2014-09-04
04-09-2014
94%
Prazo Pagamento
Pronto Pagamer
Base Incidência IVA
227.64
Taxas IVA
Taxa
Base
Valor
Taxa 1
0.23
227.64
52.36
Taxa 2
505207583 : João Brito E Cunha, Lda
QUINTA DE S. JOSÉ
RECIBO DE CLIENTE
Nº Documento: 14192
Data: 2014-09-04
ORIGINAL
João Brito e Cunha, Lda
Rua Augusto César, 99
5000-591 Vila Real
Telephone: 259 325 147
Fax: 259 325 147
Email: joaobritoecunha@quintasjose.com
URL: www.quintasjose.com
DevScope, SA
Rua Passos Manuel 223, 4º
4000-385 PORTO
Nº DOC. BANCO
BANCO
DOCUMENTOS A QUE SE REFERE ESTE DOCUMENTO
TIPO DOCUMENTO
Nº DOC.
FACTURA
14190
Page Thumbnails
Page 1
DIGITS
smartdoc-segnet-model-digits6
Test One
Login
Info
About
Clone Job
Delete Job
Generic Image Model
Job Status Done
Initialized at 12:55:58 AM (1 second)
Running at 12:55:59 AM (3 seconds)
Done at 12:56:02 AM
(Total - 4 seconds)
Infer Model Done
Source image
Inference visualization
devscope

devscope

Harvard Business Review

DATA

Getting Value from Machine Learning Isn't About Fancier Algorithms — It's About Making It Easier to Use

by Ben Schreck, Max Kanter, Kalyan Veeramachaneni, Sanjeev Vohra, and Rajendra Prasad

MARCH 06, 2018

Increasing ROI with AutoML

- Productivity (less time/less errors/more learning)
- Discovery
- Democratization/Massification
- Better Quantify Uncertainty
- Optimization (better results)
- -> Free more time to higher level tasks!

Tiny AutoML R&D Tip...

general | OpenML Sla | The Hitchhiker's Guid | Disciplina: Business Ir | Microsoft Azure Mac | TweetDeck | google 1/168 | Earl | Smarkio - Sales, Mark »

automl

All | News | Images | Videos | Books | More | Settings | Tools

About 402,000 results (0.52 seconds)

Cloud AutoML Vision | Google Cloud Platform | google.com

Ad cloud.google.com/

Create your own custom vision model with minimum Machine Learning skills needed.
Built For Developers · Highly Scalable · $300 Free Trial · Access From Anywhere · Massive Scale
Services: Integrated Services, Auto-Scaling, Traffic Splitting, Scalable App Storage, Memcache, Task ...

Get in touch

Contact us today to discuss how GCP could help your business.

Cloud AutoML - Custom Machine Learning

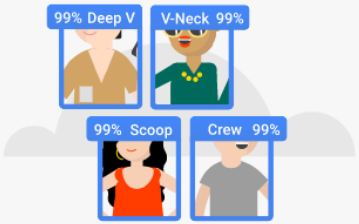
https://cloud.google.com/automl/

Cloud **AutoML** is a suite of Machine Learning products that learning expertise to train high quality models by leveraging and Neural Architecture Search technology. **AutoML** Vision simple, secure and ...

AutoML – Machine Learning for Automated

www.ml4aad.org/automl/

News. We will hold an **AutoML** workshops at the Federated

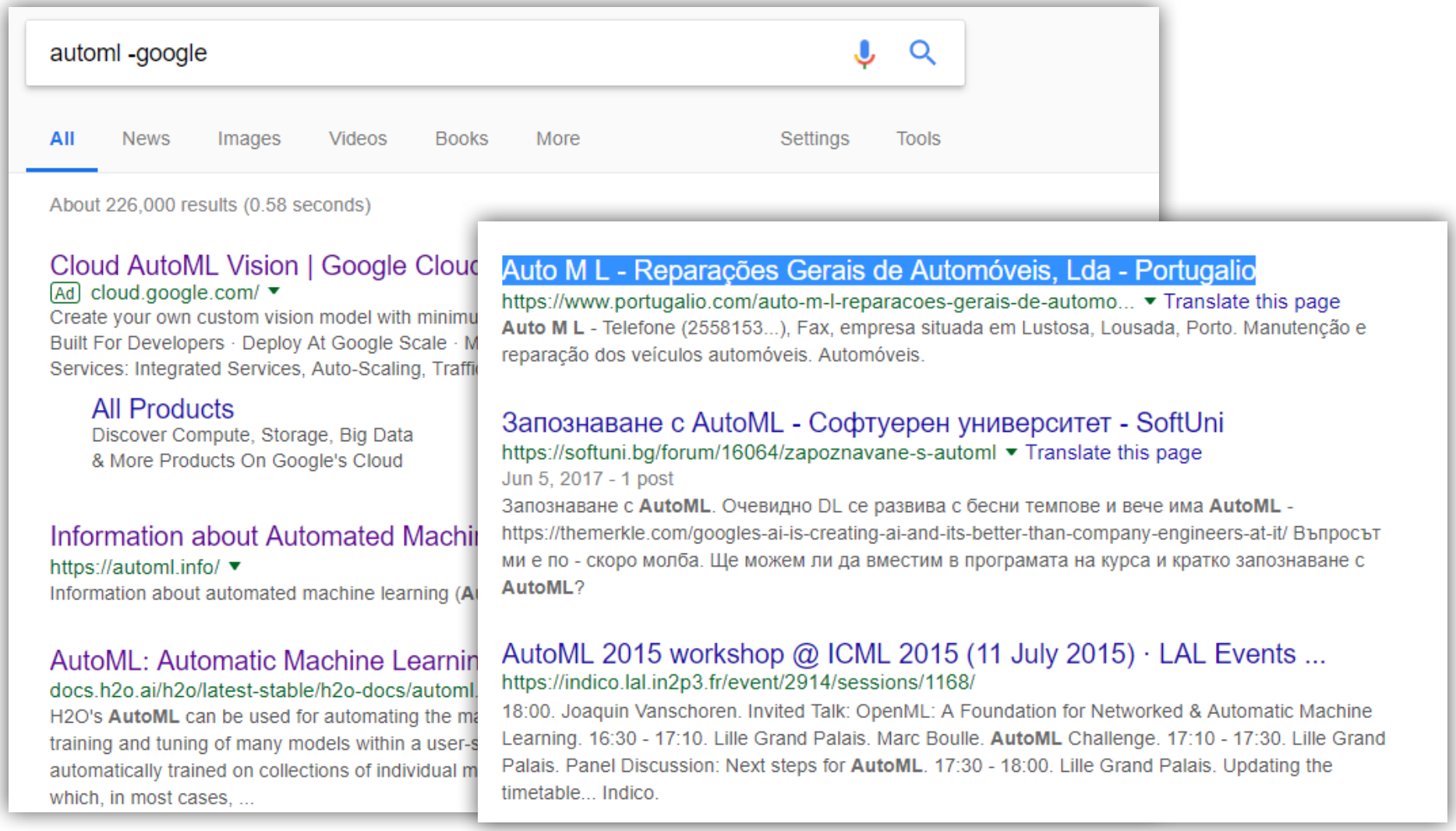


Easily train custom vision models

With Cloud AutoML, you can bring your training data to create your own custom vision model with minimum Machine Learning skills required. Start with as little as a few dozen photographic samples and Cloud AutoML will do the rest.

State-of-the-art performance

Use Cloud AutoML to leverage Google's proprietary image recognition technology. The AutoML technology achieved the state of the art performance on popular public datasets (CIFAR and ImageNet). This research is now available to you.


automl -google
All
News
Images
Videos
Books
More
Settings
Tools
About 226,000 results (0.58 seconds)
Cloud AutoML Vision | Google Cloud
Ad cloud.google.com/
Create your own custom vision model with minimu
Built For Developers · Deploy At Google Scale · M
Services: Integrated Services, Auto-Scaling, Traffi
All Products
Discover Compute, Storage, Big Data
& More Products On Google's Cloud
Information about Automated Machi
https://automl.info/
Information about automated machine learning (A
AutoML: Automatic Machine Learnin
docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.
H2O's AutoML can be used for automating the ma
training and tuning of many models within a user-s
automatically trained on collections of individual m
which, in most cases, ...
Auto M L - Reparações Gerais de Automóveis, Lda - Portugalio
https://www.portugalio.com/auto-m-l-reparacoes-gerais-de-automo... Translate this page
Auto M L - Telefone (2558153...), Fax, empresa situada em Lustosa, Lousada, Porto. Manutenção e reparação dos veículos automóveis. Automóveis.
Запознаване с AutoML - Софтуерен университет - SoftUni
https://softuni.bg/forum/16064/zapoznavane-s-automl Translate this page
Jun 5, 2017 - 1 post
Запознаване с AutoML. Очевидно DL се развива с бесни темпове и вече има AutoML - https://themerkle.com/googles-ai-is-creating-ai-and-its-better-than-company-engineers-at-it/ Въпросът ми е по - скоро молба. Ще можем ли да вместим в програмата на курса и кратко запознаване с AutoML?
AutoML 2015 workshop @ ICML 2015 (11 July 2015) · LAL Events ...
https://indico.lal.in2p3.fr/event/2914/sessions/1168/
18:00. Joaquin Vanschoren. Invited Talk: OpenML: A Foundation for Networked & Automatic Machine Learning. 16:30 - 17:10. Lille Grand Palais. Marc Boulle. AutoML Challenge. 17:10 - 17:30. Lille Grand Palais. Panel Discussion: Next steps for AutoML. 17:30 - 18:00. Lille Grand Palais. Updating the timetable... Indico.

Automated Machine Learning (in this session)

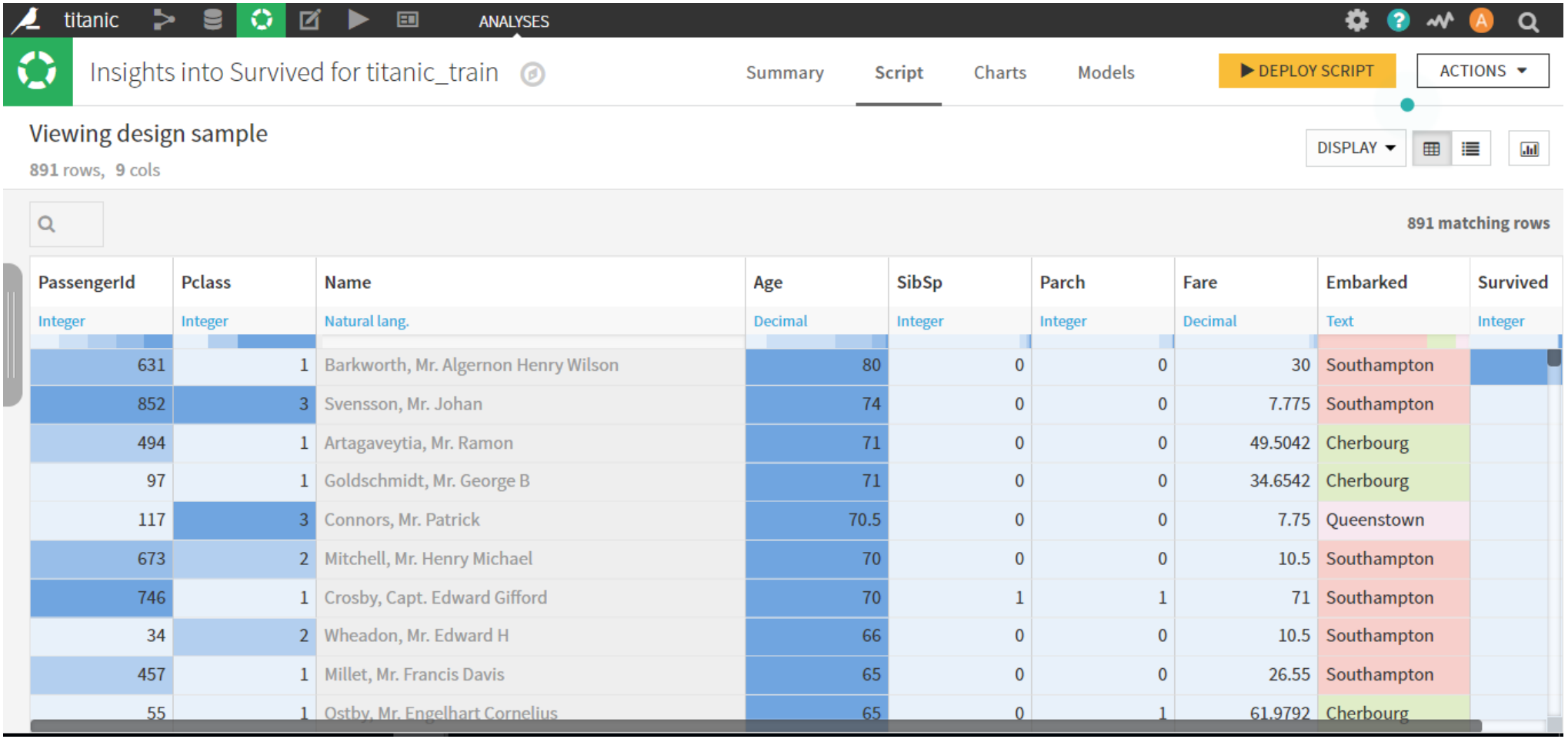
- (for Data Scientists/ML Experts)
- Common tasks, Repetitive tasks (free us)
- Search tasks/ Even “Reasoning” (augment us)

- *Ps-Not Automated Data Science*

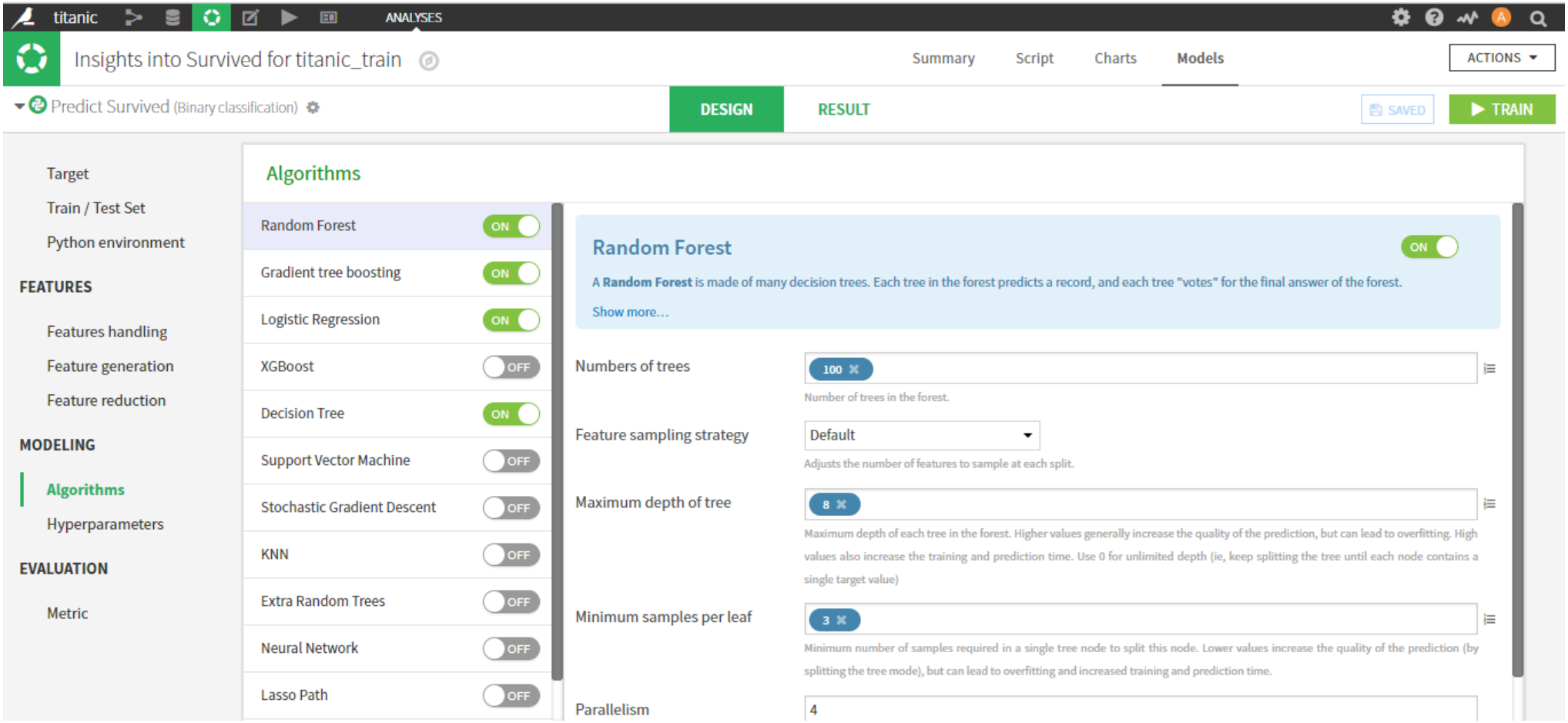
A few shades of AutoML...

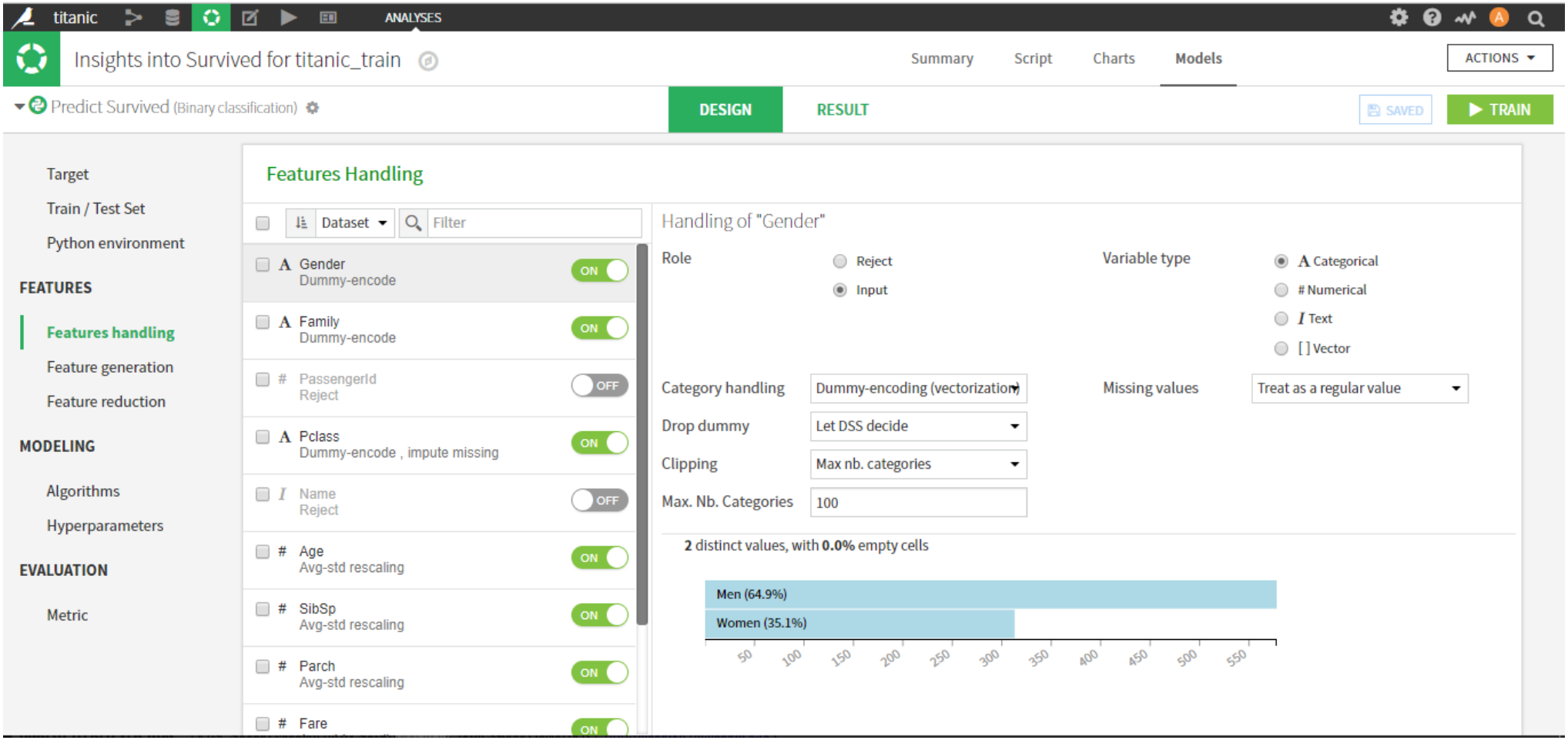
Machine Learning GUIs - Productivity

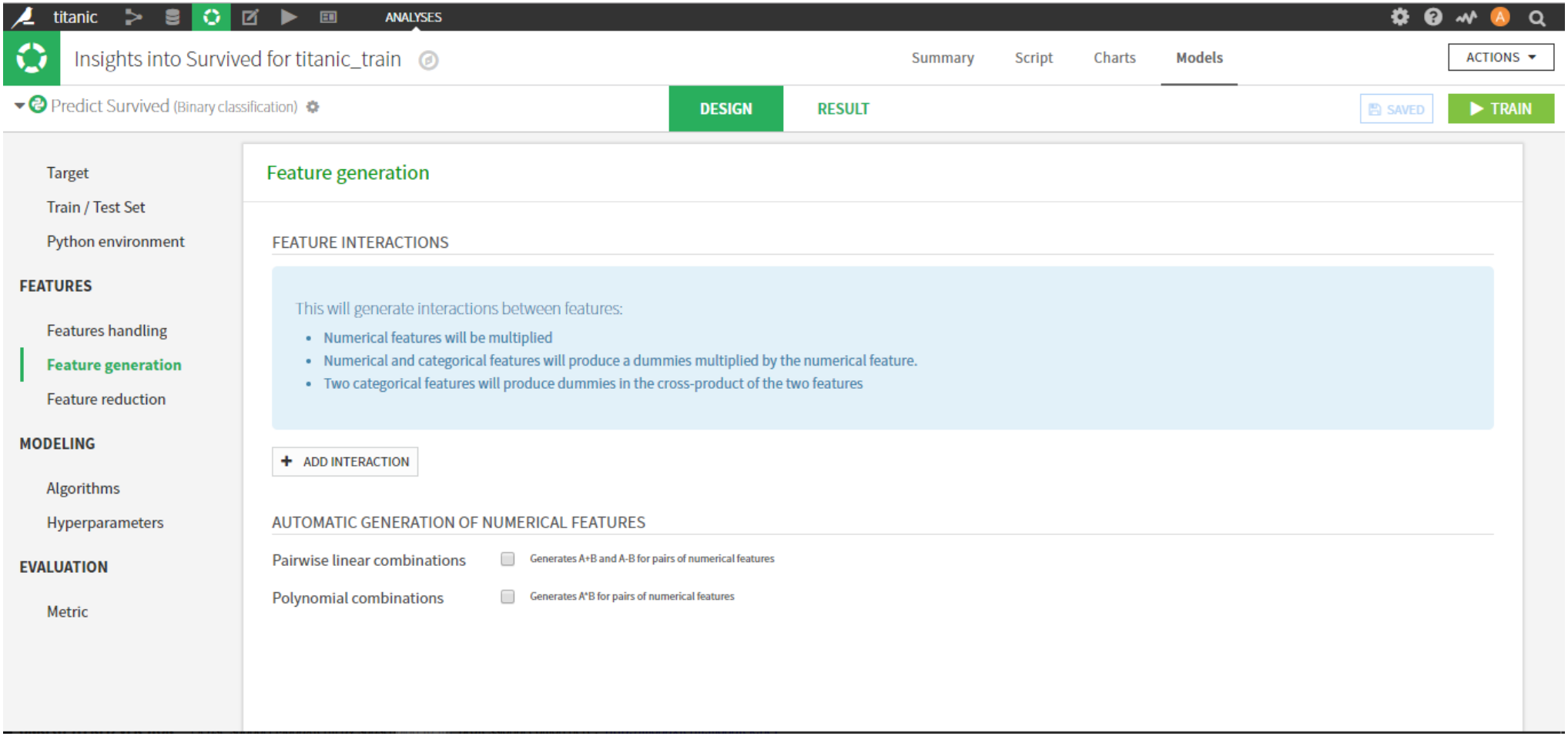
```
>docker run -it -p 10000:10000 dataiku/dss
```

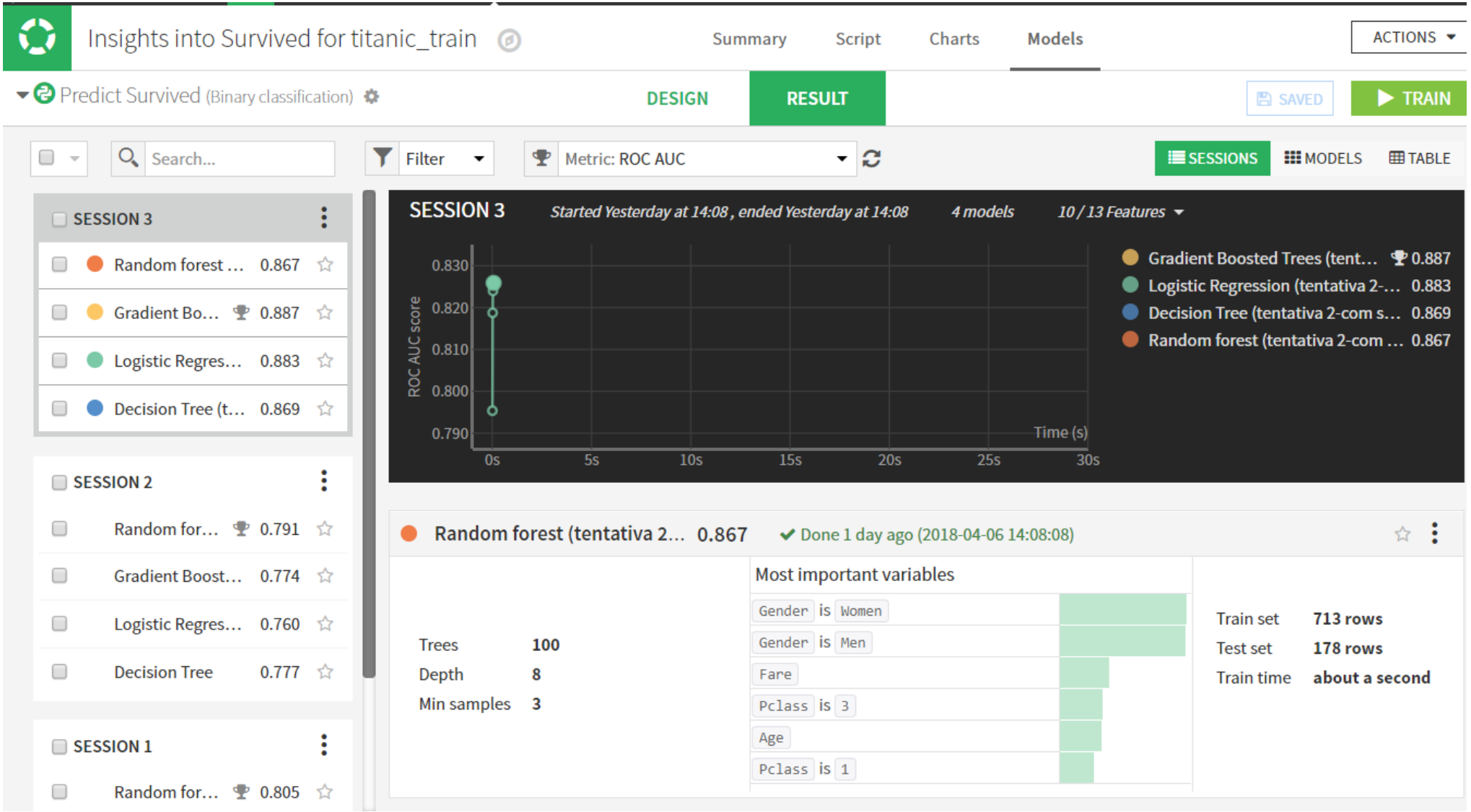


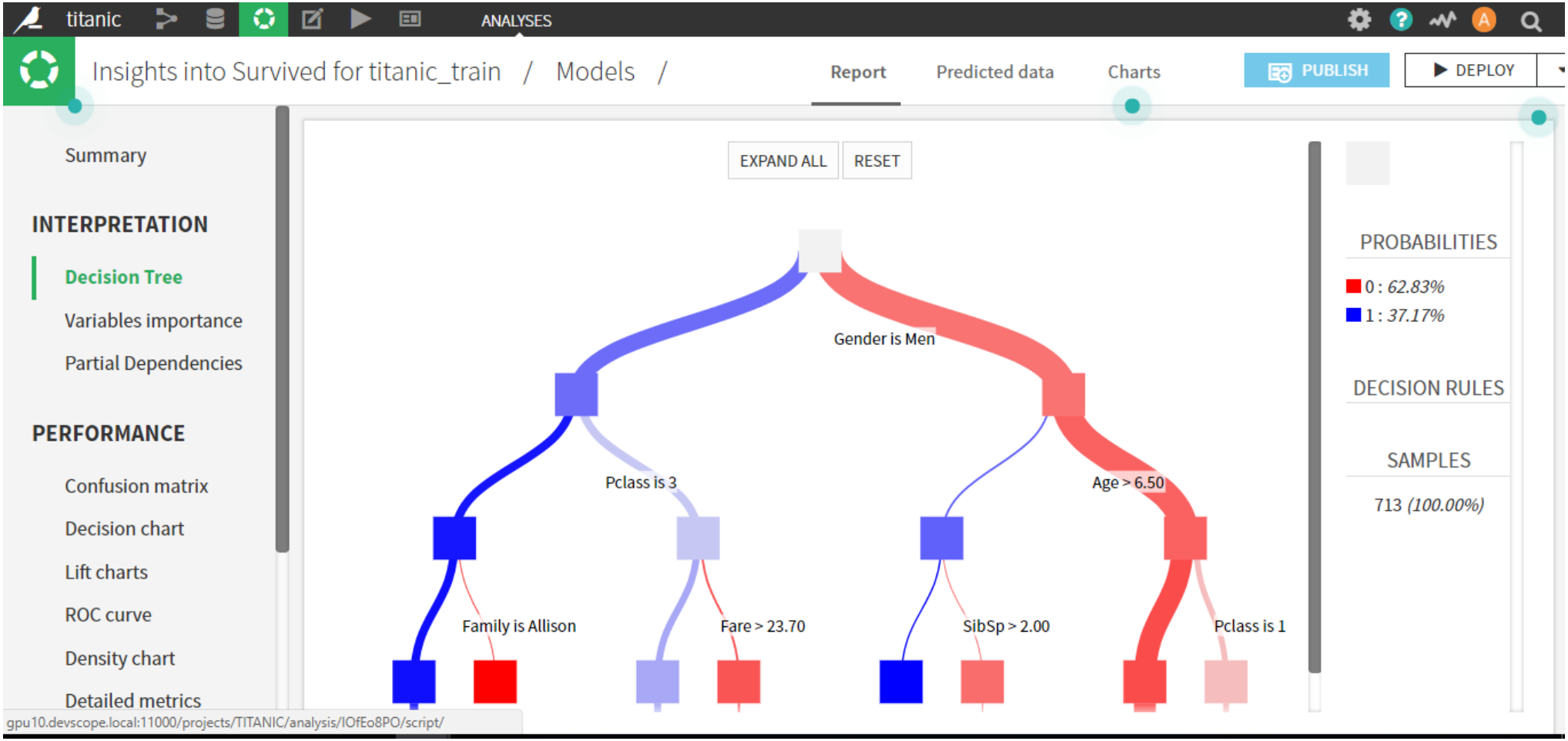
PassengerId	Pclass	Name	Age	SibSp	Parch	Fare	Embarked	Survived
Integer	Integer	Natural lang.	Decimal	Integer	Integer	Decimal	Text	Integer
631	1	Barkworth, Mr. Algernon Henry Wilson	80	0	0	30	Southampton	
852	3	Svensson, Mr. Johan	74	0	0	7.775	Southampton	
494	1	Artagaveytia, Mr. Ramon	71	0	0	49.5042	Cherbourg	
97	1	Goldschmidt, Mr. George B	71	0	0	34.6542	Cherbourg	
117	3	Connors, Mr. Patrick	70.5	0	0	7.75	Queenstown	
673	2	Mitchell, Mr. Henry Michael	70	0	0	10.5	Southampton	
746	1	Crosby, Capt. Edward Gifford	70	1	1	71	Southampton	
34	2	Wheadon, Mr. Edward H	66	0	0	10.5	Southampton	
457	1	Millet, Mr. Francis Davis	65	0	0	26.55	Southampton	
55	1	Ostby, Mr. Engelhart Cornelius	65	0	1	61.9792	Cherbourg	

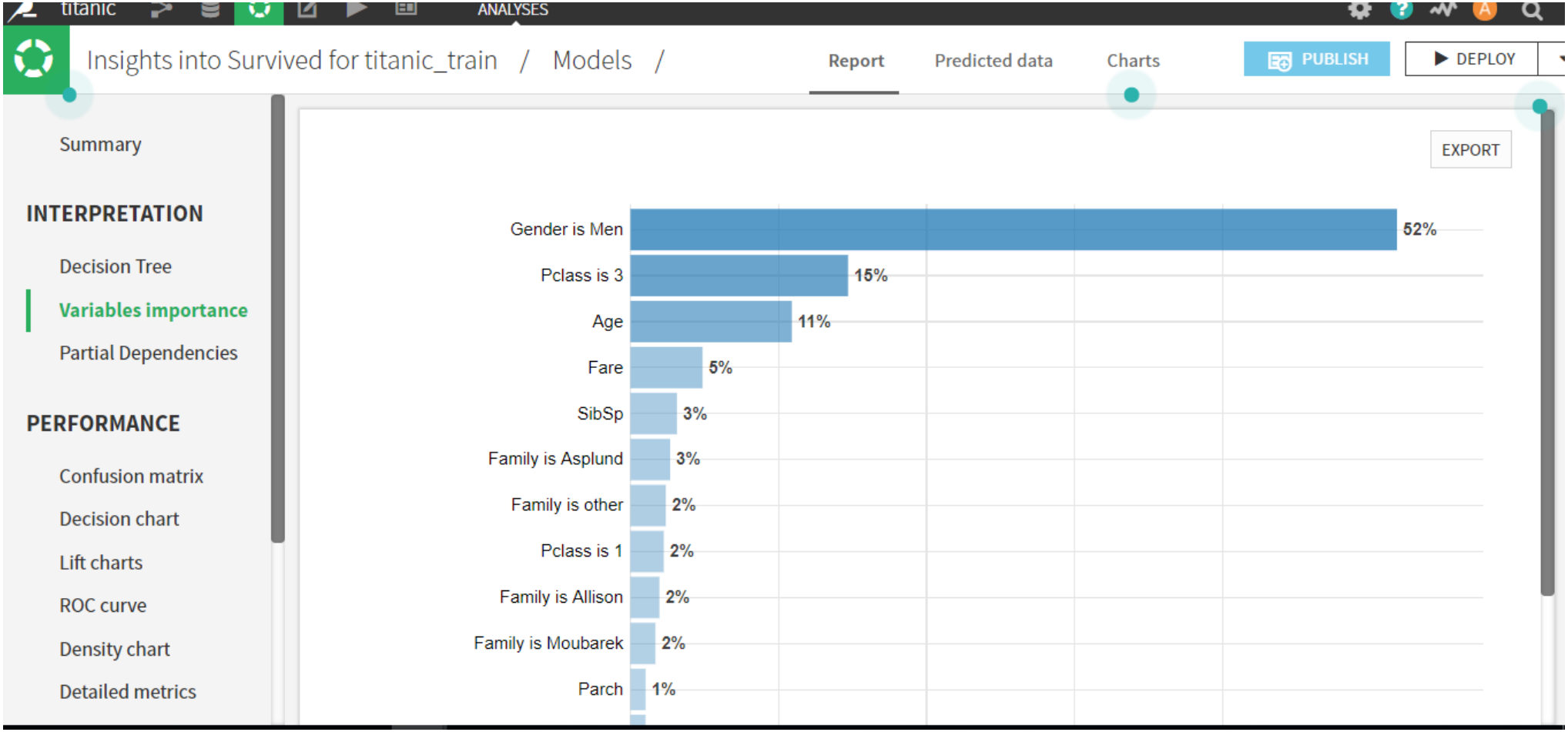

titanic
ANALYSES
Insights into Survived for titanic_train
Summary
Script
Charts
Models
ACTIONS
Predict Survived (Binary classification)
DESIGN
RESULT
SAVED
TRAIN
Target
Train / Test Set
Python environment
FEATURES
Features handling
Feature generation
Feature reduction
MODELING
Algorithms
Hyperparameters
EVALUATION
Metric
Algorithms
Random Forest ON
Gradient tree boosting ON
Logistic Regression ON
XGBoost OFF
Decision Tree ON
Support Vector Machine OFF
Stochastic Gradient Descent OFF
KNN OFF
Extra Random Trees OFF
Neural Network OFF
Lasso Path OFF
Random Forest
ON
A Random Forest is made of many decision trees. Each tree in the forest predicts a record, and each tree "votes" for the final answer of the forest.
Show more…
Numbers of trees
100
Number of trees in the forest.
Feature sampling strategy
Default
Adjusts the number of features to sample at each split.
Maximum depth of tree
8
Maximum depth of each tree in the forest. Higher values generally increase the quality of the prediction, but can lead to overfitting. High values also increase the training and prediction time. Use 0 for unlimited depth (ie, keep splitting the tree until each node contains a single target value)
Minimum samples per leaf
3
Minimum number of samples required in a single tree node to split this node. Lower values increase the quality of the prediction (by splitting the tree mode), but can lead to overfitting and increased training and prediction time.
Parallelism
4

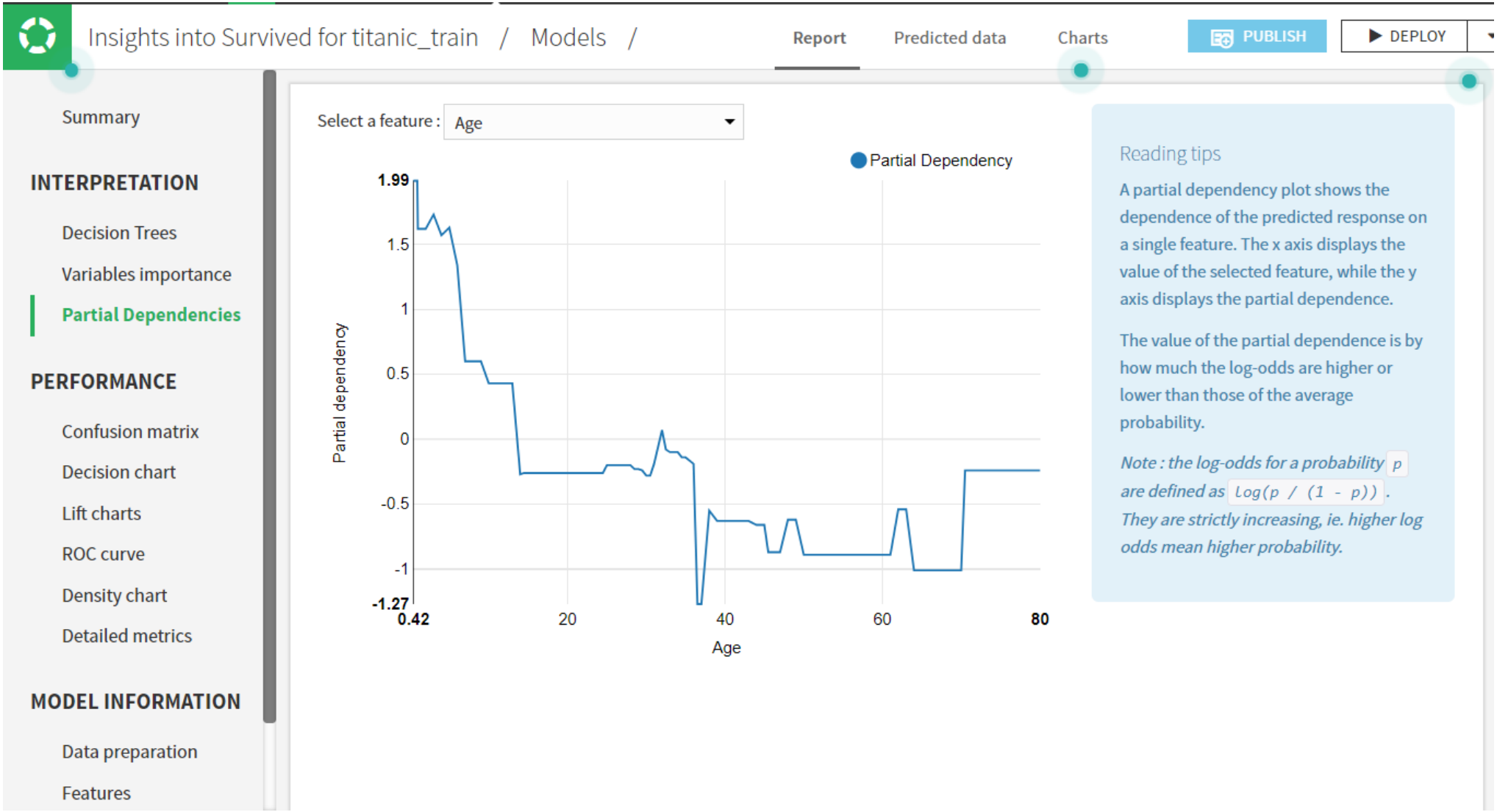

titanic
ANALYSES
Insights into Survived for titanic_train
Summary
Script
Charts
Models
ACTIONS
Predict Survived (Binary classification)
DESIGN
RESULT
SAVED
TRAIN
Target
Train / Test Set
Python environment
FEATURES
Features handling
Feature generation
Feature reduction
MODELING
Algorithms
Hyperparameters
EVALUATION
Metric
Features Handling
Dataset
Filter
Gender
Dummy-encode
Family
Dummy-encode
PassengerId
Reject
Pclass
Dummy-encode , impute missing
Name
Reject
Age
Avg-std rescaling
SibSp
Avg-std rescaling
Parch
Avg-std rescaling
Fare
ON
OFF
Handling of "Gender"
Role
Reject
Input
Variable type
Categorical
Numerical
Text
Vector
Category handling
Dummy-encoding (vectorization)
Missing values
Treat as a regular value
Drop dummy
Let DSS decide
Clipping
Max nb. categories
Max. Nb. Categories
100
2 distinct values, with 0.0% empty cells
Men (64.9%)
Women (35.1%)
50
100
150
200
250
300
350
400
450
500
550

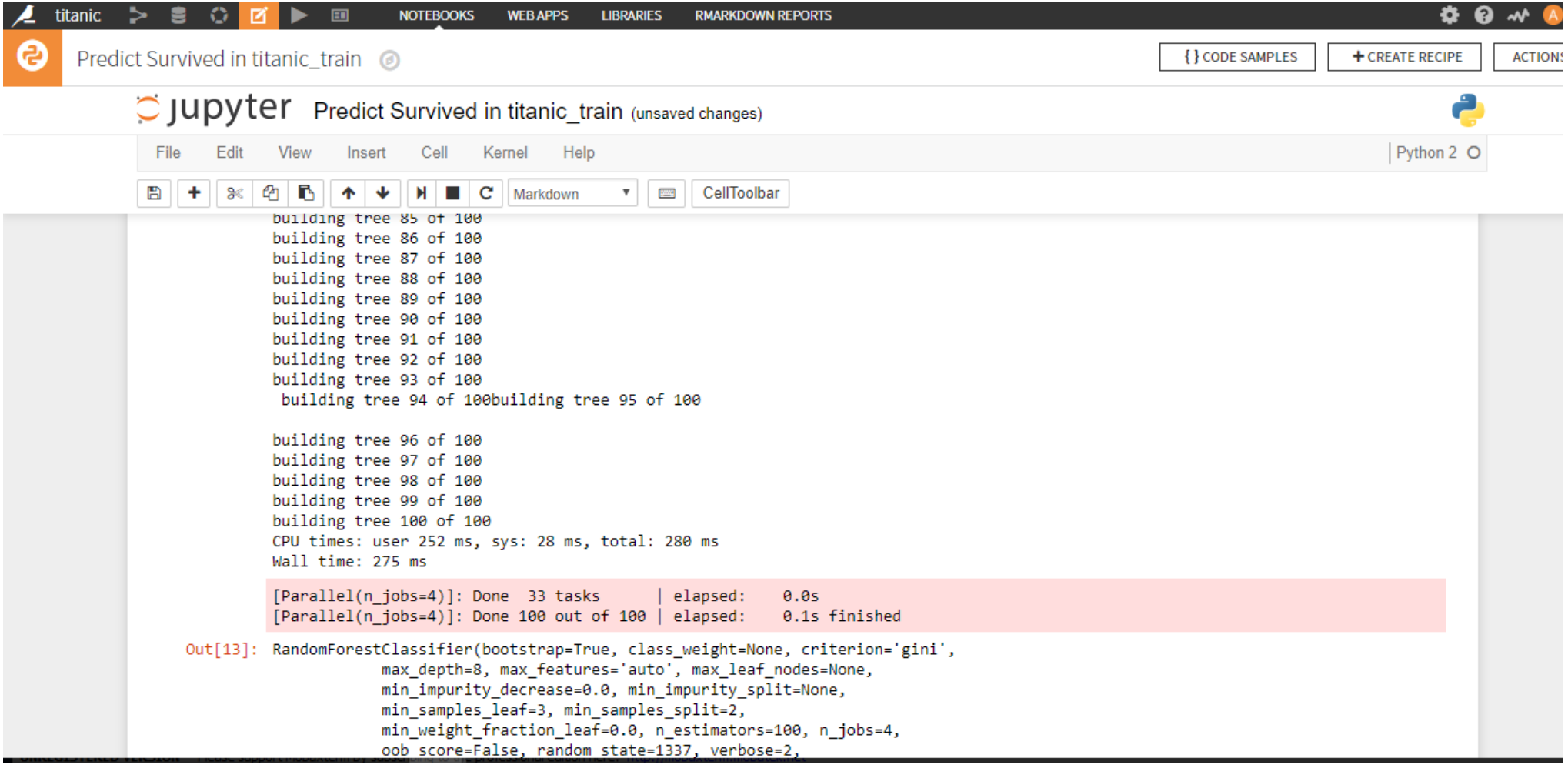

titanic
ANALYSES
Insights into Survived for titanic_train
Summary
Script
Charts
Models
ACTIONS
Predict Survived (Binary classification)
DESIGN
RESULT
SAVED
TRAIN
Target
Train / Test Set
Python environment
FEATURES
Features handling
Feature generation
Feature reduction
MODELING
Algorithms
Hyperparameters
EVALUATION
Metric
Feature generation
FEATURE INTERACTIONS
This will generate interactions between features:
Numerical features will be multiplied
Numerical and categorical features will produce a dummies multiplied by the numerical feature.
Two categorical features will produce dummies in the cross-product of the two features
ADD INTERACTION
AUTOMATIC GENERATION OF NUMERICAL FEATURES
Pairwise linear combinations
Generates A+B and A-B for pairs of numerical features
Polynomial combinations
Generates A*B for pairs of numerical features


Insights into Survived for titanic_train
Summary
Script
Charts
Models
ACTIONS
Predict Survived (Binary classification)
DESIGN
RESULT
SAVED
TRAIN
Search...
Filter
Metric: ROC AUC
SESSIONS
MODELS
TABLE
SESSION 3
Random forest … 0.867
Gradient Bo… 0.887
Logistic Regres… 0.883
Decision Tree (t… 0.869
SESSION 2
Random for… 0.791
Gradient Boost… 0.774
Logistic Regres… 0.760
Decision Tree 0.777
SESSION 1
Random for… 0.805
SESSION 3
Started Yesterday at 14:08 , ended Yesterday at 14:08
4 models
10 / 13 Features
ROC AUC score
0.830
0.820
0.810
0.800
0.790
Time (s)
0s
5s
10s
15s
20s
25s
30s
Gradient Boosted Trees (tent… 0.887
Logistic Regression (tentativa 2-… 0.883
Decision Tree (tentativa 2-com s… 0.869
Random forest (tentativa 2-com … 0.867
Random forest (tentativa 2… 0.867
Done 1 day ago (2018-04-06 14:08:08)
Trees 100
Depth 8
Min samples 3
Most important variables
Gender is Women
Gender is Men
Fare
Pclass is 3
Age
Pclass is 1
Train set 713 rows
Test set 178 rows
Train time about a second


titanic
ANALYSES
Insights into Survived for titanic_train / Models /
Report
Predicted data
Charts
PUBLISH
DEPLOY
Summary
INTERPRETATION
Decision Tree
Variables importance
Partial Dependencies
PERFORMANCE
Confusion matrix
Decision chart
Lift charts
ROC curve
Density chart
Detailed metrics
EXPAND ALL
RESET
Gender is Men
Pclass is 3
Age > 6.50
Family is Allison
Fare > 23.70
SibSp > 2.00
Pclass is 1
PROBABILITIES
0 : 62.83%
1 : 37.17%
DECISION RULES
SAMPLES
713 (100.00%)
gpu10.devscope.local:11000/projects/TITANIC/analysis/IOfEo8PO/script/


titanic
ANALYSES
Insights into Survived for titanic_train / Models /
Report
Predicted data
Charts
PUBLISH
DEPLOY
Summary
INTERPRETATION
Decision Tree
Variables importance
Partial Dependencies
PERFORMANCE
Confusion matrix
Decision chart
Lift charts
ROC curve
Density chart
Detailed metrics
EXPORT
Gender is Men 52%
Pclass is 3 15%
Age 11%
Fare 5%
SibSp 3%
Family is Asplund 3%
Family is other 2%
Pclass is 1 2%
Family is Allison 2%
Family is Moubarek 2%
Parch 1%


Insights into Survived for titanic_train / Models /
Report
Predicted data
Charts
PUBLISH
DEPLOY
Summary
INTERPRETATION
Decision Trees
Variables importance
Partial Dependencies
PERFORMANCE
Confusion matrix
Decision chart
Lift charts
ROC curve
Density chart
Detailed metrics
MODEL INFORMATION
Data preparation
Features
Select a feature : Age
Partial Dependency
1.99
1.5
1
0.5
0
-0.5
-1
-1.27
Partial dependency
0.42
20
40
60
80
Age
Reading tips
A partial dependency plot shows the dependence of the predicted response on a single feature. The x axis displays the value of the selected feature, while the y axis displays the partial dependence.
The value of the partial dependence is by how much the log-odds are higher or lower than those of the average probability.
Note : the log-odds for a probability p are defined as log(p / (1 - p)). They are strictly increasing, ie. higher log odds mean higher probability.


titanic
NOTEBOOKS
WEB APPS
LIBRARIES
RMARKDOWN REPORTS
Predict Survived in titanic_train
{ } CODE SAMPLES
+ CREATE RECIPE
ACTIONS
jupyter
Predict Survived in titanic_train (unsaved changes)
File
Edit
View
Insert
Cell
Kernel
Help
Python 2
Markdown
CellToolbar
building tree 85 of 100
building tree 86 of 100
building tree 87 of 100
building tree 88 of 100
building tree 89 of 100
building tree 90 of 100
building tree 91 of 100
building tree 92 of 100
building tree 93 of 100
building tree 94 of 100building tree 95 of 100
building tree 96 of 100
building tree 97 of 100
building tree 98 of 100
building tree 99 of 100
building tree 100 of 100
CPU times: user 252 ms, sys: 28 ms, total: 280 ms
Wall time: 275 ms
[Parallel(n_jobs=4)]: Done 33 tasks | elapsed: 0.0s
[Parallel(n_jobs=4)]: Done 100 out of 100 | elapsed: 0.1s finished
Out[13]: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
max_depth=8, max_features='auto', max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=3, min_samples_split=2,
min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=4,
oob_score=False, random_state=1337, verbose=2,

Model/Pipeline Selection, Hyper-Parameter Tuning

H2O *AutoML (R/Python)*

r python

```r
library(h2o)

h2o.init()

# Import a sample binary outcome train/test set into H2O
train <- h2o.importFile("https://s3.amazonaws.com/erin-data/higgs
test <- h2o.importFile("https://s3.amazonaws.com/erin-data/higgs/

# Identify predictors and response
y <- "response"
x <- setdiff(names(train), y)

# For binary classification, response should be a factor
train[,y] <- as.factor(train[,y])
test[,y] <- as.factor(test[,y])

aml <- h2o.automl(x = x, y = y,
                  training_frame = train,
                  max_runtime_secs = 30)
```

r python

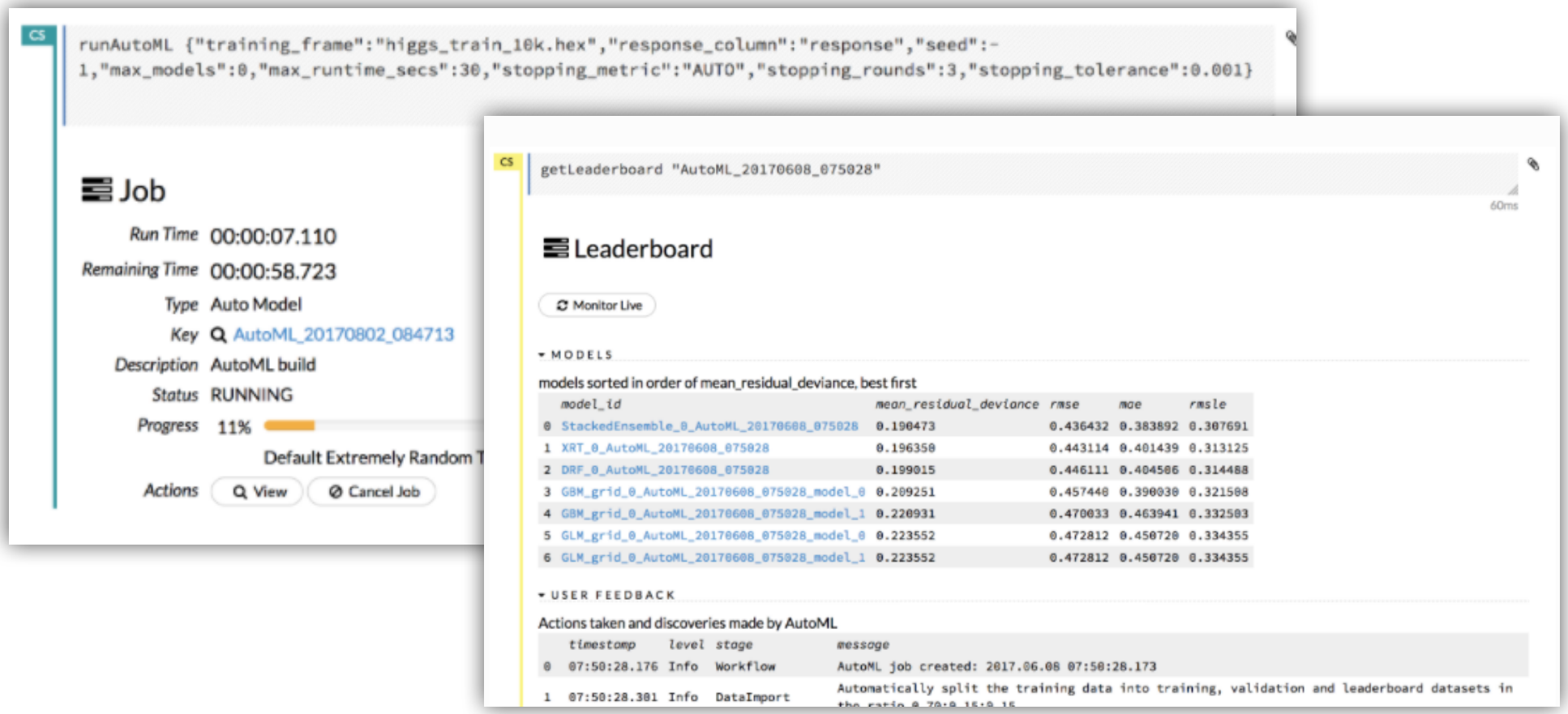
```python
import h2o
from h2o.automl import H2OAutoML

h2o.init()

# Import a sample binary outcome train/test set into H2O
train = h2o.import_file("https://s3.amazonaws.com/erin-data/higgs/higgs_train_10k.csv")
test = h2o.import_file("https://s3.amazonaws.com/erin-data/higgs/higgs_test_5k.csv")

# Identify predictors and response
x = train.columns
y = "response"
x.remove(y)

# For binary classification, response should be a factor
train[y] = train[y].asfactor()
test[y] = test[y].asfactor()

# Run AutoML for 30 seconds
aml = H2OAutoML(max_runtime_secs = 30)
aml.train(x = x, y = y,
```

H2O *AutoML (R/Python)*

```
# View the AutoML Leaderboard
lb <- aml@leaderboard
lb

#                                                  model_id      auc  logloss
```

model_id	auc	logloss
StackedEnsemble_AllModels_0_AutoML_20171121_012135	0.788321	0.554019
StackedEnsemble_BestOfFamily_0_AutoML_20171121_012135	0.783099	0.559286
GBM_grid_0_AutoML_20171121_012135_model_1	0.780554	0.560248
GBM_grid_0_AutoML_20171121_012135_model_0	0.779713	0.562142
GBM_grid_0_AutoML_20171121_012135_model_2	0.776206	0.564970
GBM_grid_0_AutoML_20171121_012135_model_3	0.771026	0.570270
DRF_0_AutoML_20171121_012135	0.734653	0.601520
XRT_0_AutoML_20171121_012135	0.730457	0.611706
GBM_grid_0_AutoML_20171121_012135_model_4	0.727098	0.666513
GLM_grid_0_AutoML_20171121_012135_model_0	0.685211	0.635138

H2O Flow UI / Leaderboard



Tpot (python)

- *(genetic search)*

```
from tpot import TPOTClassifier
from sklearn.datasets import load_digits
from sklearn.model_selection import train_test_split

digits = load_digits()
X_train, X_test, y_train, y_test = train_test_split(digits.data, digits.target,
                                                    train_size=0.75, test_size=0.25)

tpot = TPOTClassifier(generations=5, population_size=50, verbosity=2, n_jobs=-1)
tpot.fit(X_train, y_train)
```

```
Optimization Progress:  33%|███▍       | 100/300 [01:02<09:07,  2.74s/pipeline]
Generation 1 - Current best internal CV score: 0.9644750872792087
Optimization Progress:  50%|█████      | 150/300 [01:35<05:41,  2.27s/pipeline]
Generation 2 - Current best internal CV score: 0.9681323584103183
Optimization Progress:  67%|██████▋    | 200/300 [01:59<01:59,  1.19s/pipeline]
Generation 3 - Current best internal CV score: 0.9718282518620386
Optimization Progress:  83%|████████▎  | 250/300 [02:23<00:41,  1.21pi
Generation 4 - Current best internal CV score: 0.9756538552070356

Generation 5 - Current best internal CV score: 0.9756538552070356

Best pipeline: KNeighborsClassifier(input_matrix, KNeighborsClassifier__n_neighbors=10, KNeighborsClassifier__p=DEFAU
LT, KNeighborsClassifier__weights=distance)
```

~99% accuracy on MNIST out of the box

```
print(tpot.score(X_test, y_test))
```

```
0.995555555556
```

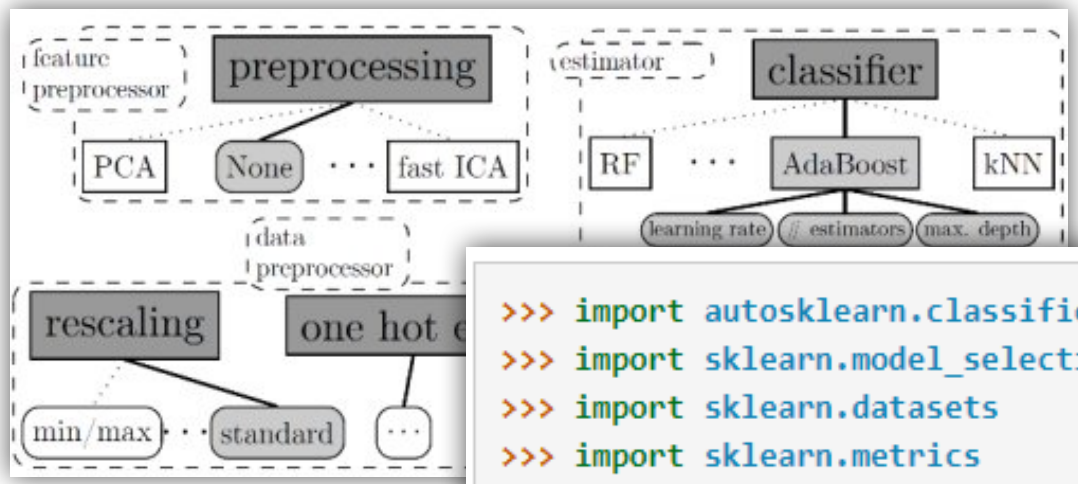
Tpot (python)

- Export pipeline

should be exported to the `tpot_mnist_pipeline.py` file and look similar to the following:

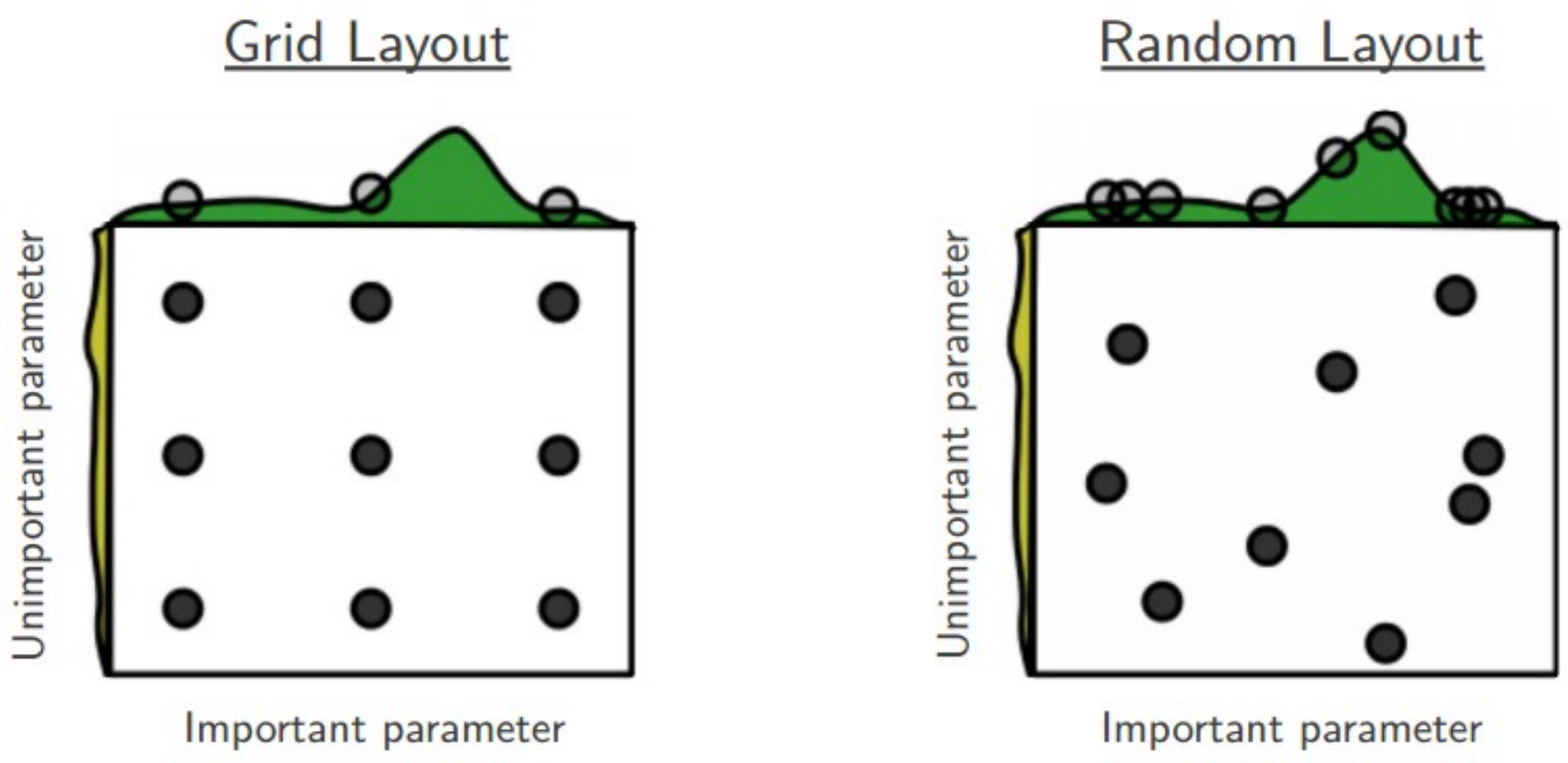
```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier

# NOTE: Make sure that the class is labeled 'target' in the data file
tpot_data = pd.read_csv('PATH/TO/DATA/FILE', sep='COLUMN_SEPARATOR', dtype=np.float64)
features = tpot_data.drop('target', axis=1).values
training_features, testing_features, training_target, testing_target = \
            train_test_split(features, tpot_data['target'].values, random_state=42)


exported_pipeline = KNeighborsClassifier(n_neighbors=6, weights="distance")

exported_pipeline.fit(training_features, training_classes)
results = exported_pipeline.predict(testing_features)
```

Auto-sklearn

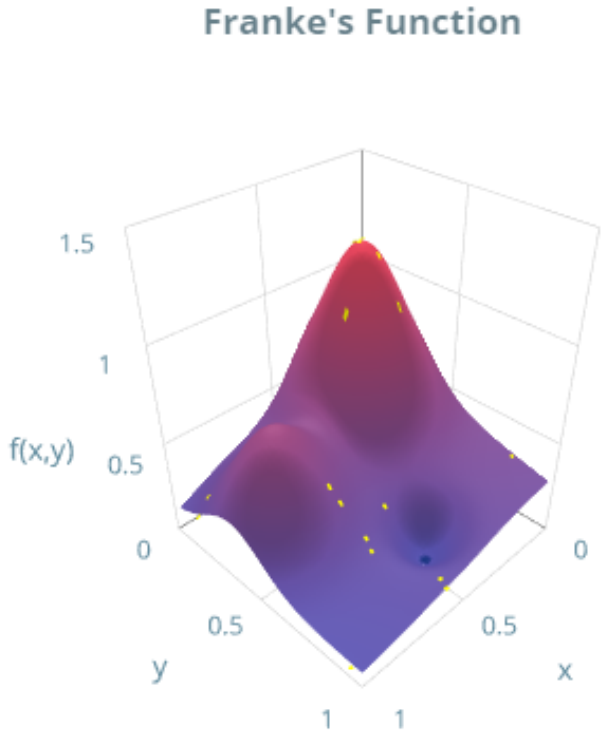


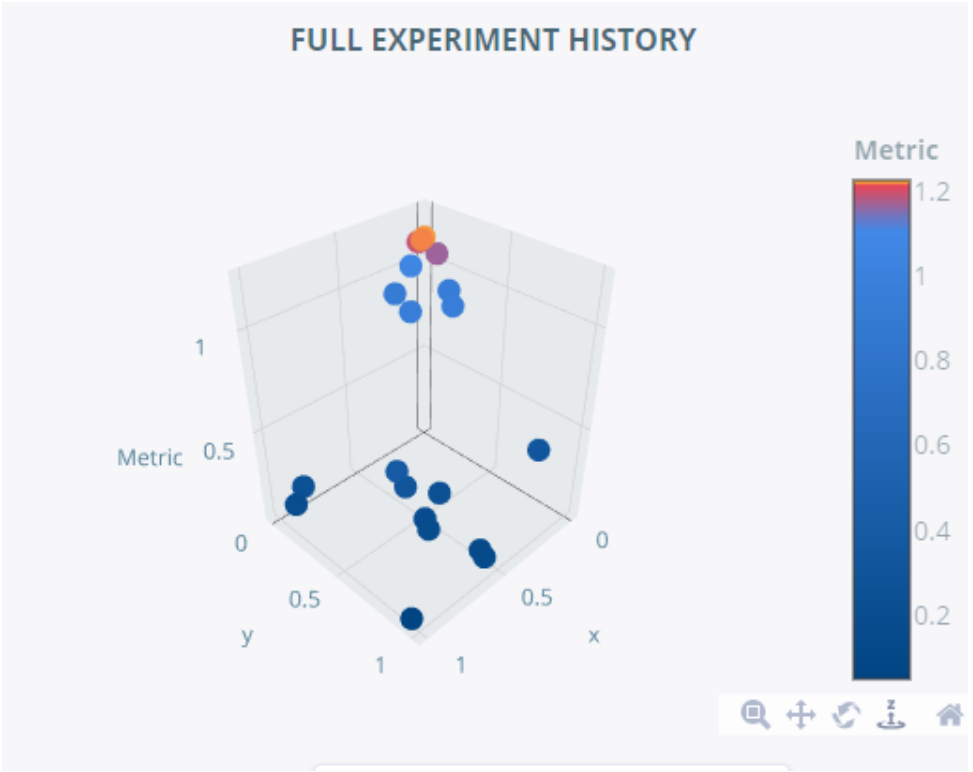

```
>>> import autosklearn.classification
>>> import sklearn.model_selection
>>> import sklearn.datasets
>>> import sklearn.metrics
>>> X, y = sklearn.datasets.load_digits(return_X_y=True)
>>> X_train, X_test, y_train, y_test = \
        sklearn.model_selection.train_test_split(X, y, random_state=1)
>>> automl = autosklearn.classification.AutoSklearnClassifier()
>>> automl.fit(X_train, y_train)
>>> y_hat = automl.predict(X_test)
>>> print("Accuracy score", sklearn.metrics.accuracy_score(y_test, y_hat))
```

Search Strategies: Random & Grid

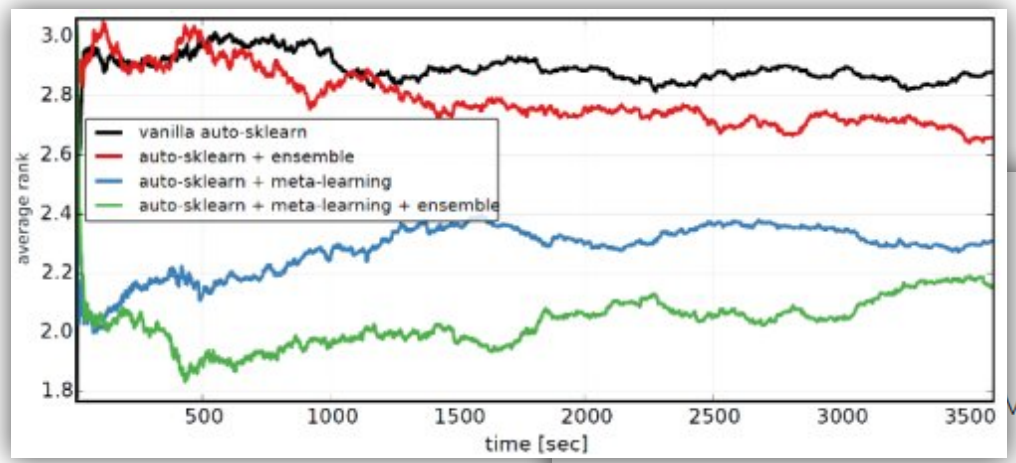


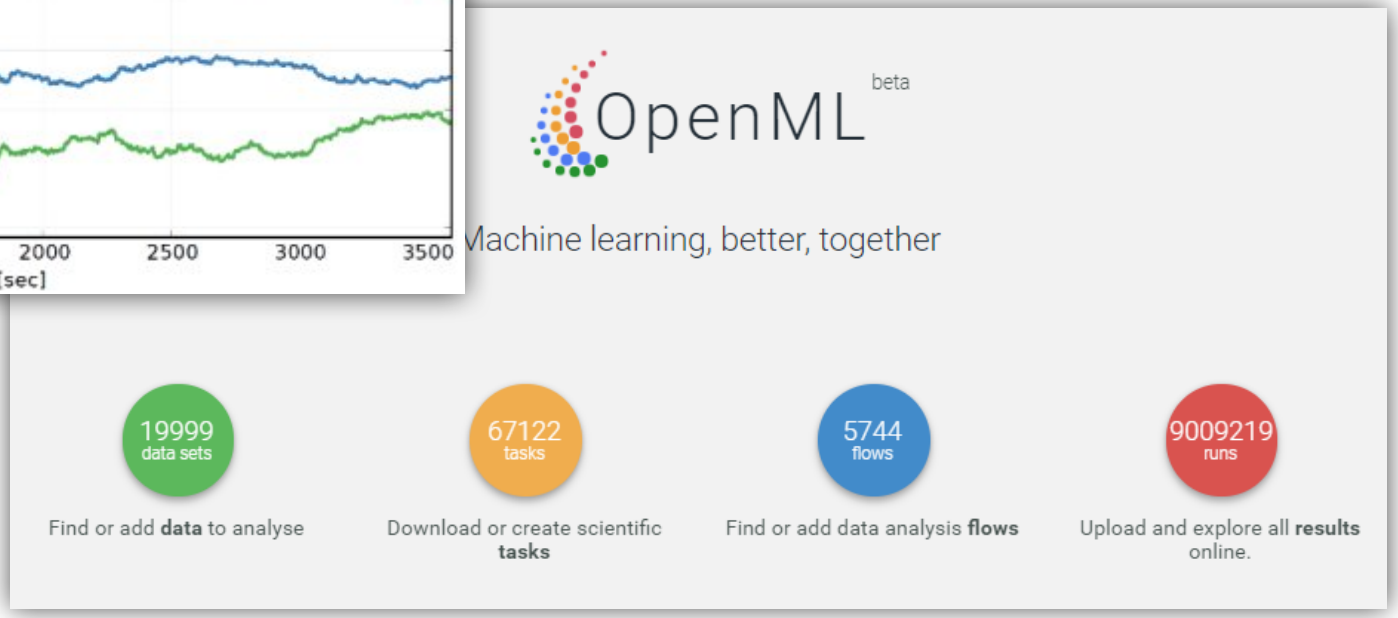
Search Strategies: Bayesian





Bayesian Search & Meta Learning – Fast Start!

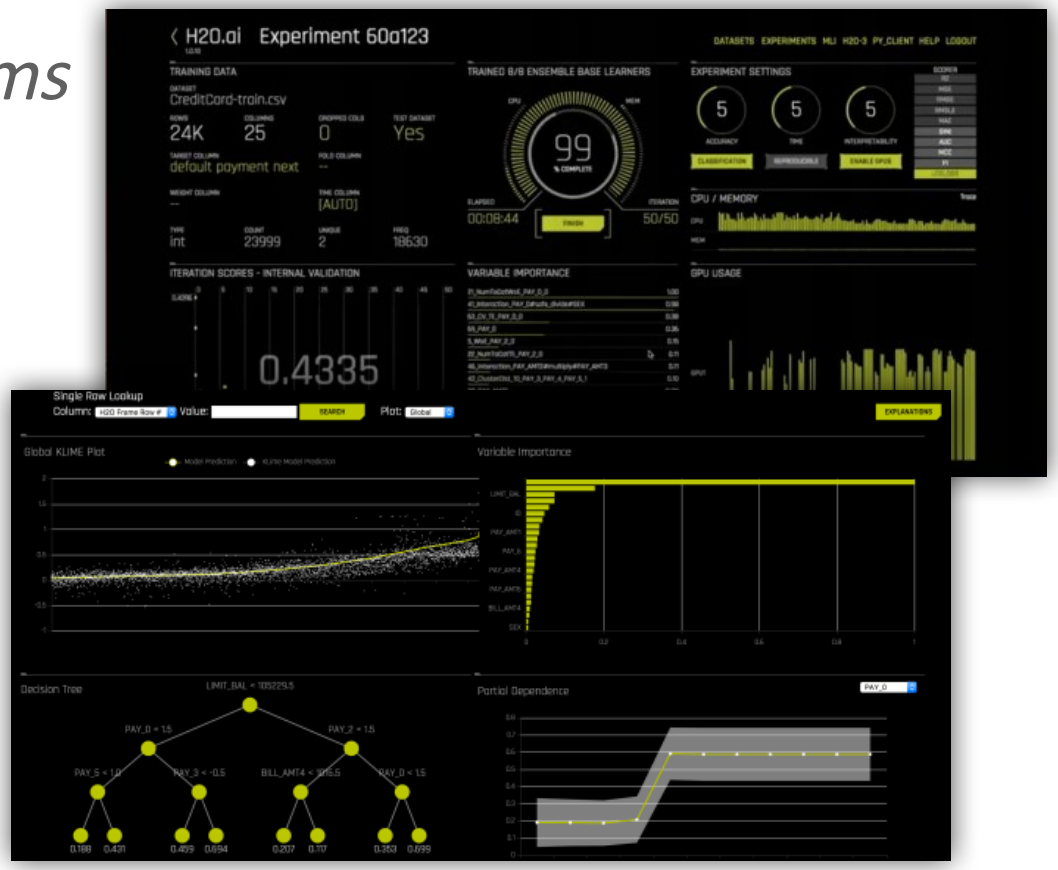




Commercial AutoML Kaggle “Grandmasters”

H2O Driverless AI (Enterprise, Not O.S.)

- *Algorithms/Hyper Params*
- *Ensembles*
- *Feature Engineering*
- *Interpretability*
- *Fast (GPUs)*



H2O.ai
0.0.1

Show All Experiments

TRAINING DATA

DATASET
creditcard.csv

ROWS	COLUMNS	DROPPED	IGNORED
24K	25	--	--

TARGET COLUMN
default payment next month

TYPE	COUNT	MEAN	STD DEV
int32	23999	0.2237	0.4167

EXPERIMENT SETTINGS

50	3	4
ITERATIONS	CV FOLDS	POPULATION

CLASSIFICATION | ONE SPLIT ONLY | AUTO SEL. COLS

SCORER
- R2
- AUC
- RMSE
- MSE
- MAE
- LOGLOSS

LAUNCH EXPERIMENT

H2O.ai

0.0.1

TRAINING DATA

DATASET

creditcard.csv

ROWS	COLUMNS	DROPPED	IGNORED
24K	25	--	--

TARGET COLUMN

default payment next month

TYPE	COUNT	MEAN	STD DEV
int32	23999	0.2237	0.4167

STATUS: COMPLETE

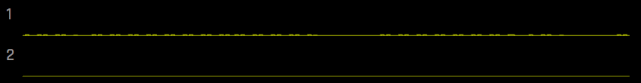
INTERPRET THIS MODEL

EXPERIMENT SETTINGS

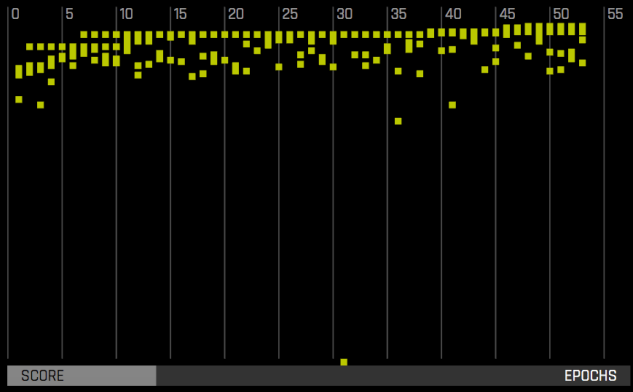
1	53	3	4
WORKERS	ITERATIONS	CV FOLDS	POPULATION

CLASSIFICATION | DETECT IDS | DROP DUPS.

GPU STATS



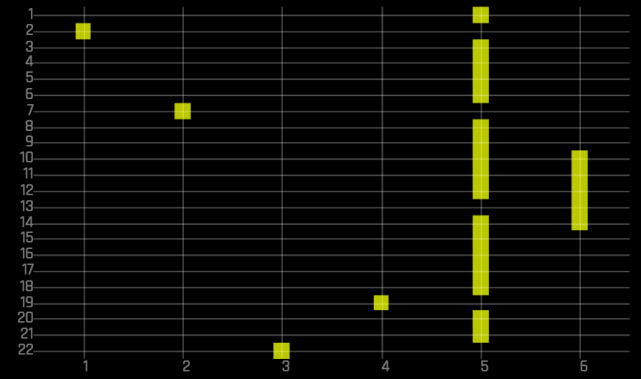
ITERATION SCORES



VARIABLE IMPORTANCE

1_PAY_0	7497.51
2_PAY_2	1020.18
5_BILL_AMT1	888.59
0_LIMIT_BAL	826.45
11_PAY_AMT2	777.38
10_PAY_AMT1	684.03
20_CV_TE_SEX_EDUCATION_0	634.89
12_PAY_AMT3	503.78
13_PAY_AMT4	499.84
6_BILL_AMT2	428.59
9_BILL_AMT5	425.90
14_PAY_AMT5	418.09
15_PAY_AMT6	409.82
3_PAY_4	326.56

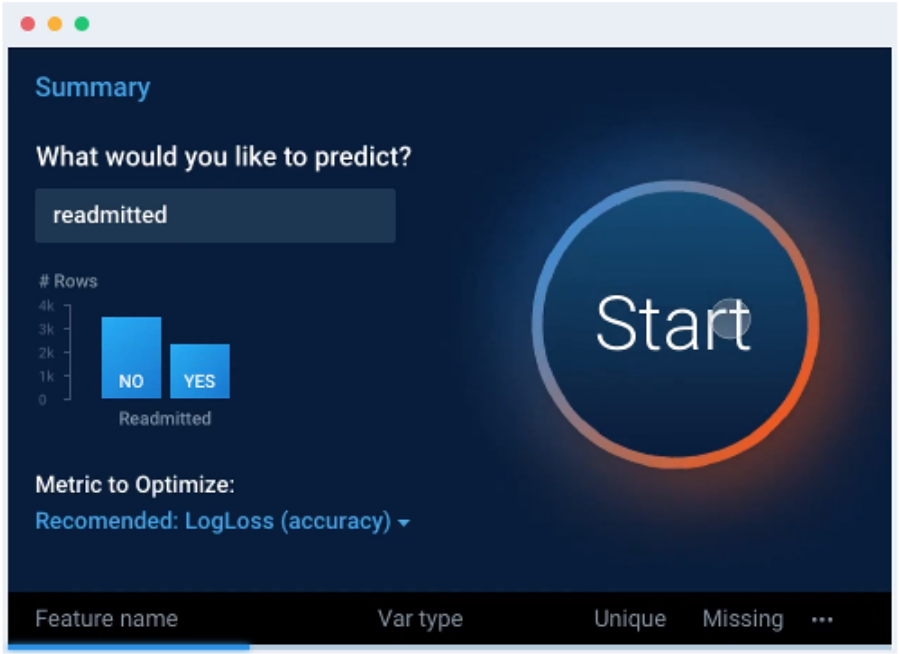
FEATURE TRANSFORMATIONS




Single Row Lookup
Column: H2O Frame Row #
Value:
SEARCH
Plot: Global
EXPLANATIONS
Global KLIME Plot
Model Prediction
KLime Model Prediction
2
1.5
1
0.5
0
-0.5
-1
Variable Importance
LIMIT_BAL
ID
PAY_AMT1
PAY_5
PAY_AMT4
PAY_AMT5
BILL_AMT4
SEX
0
0.2
0.4
0.6
0.8
1
Decision Tree
LIMIT_BAL < 105229.5
PAY_0 < 1.5
PAY_2 < 1.5
PAY_5 < 1.0
PAY_3 < -0.5
BILL_AMT4 < 1016.5
PAY_0 < 1.5
0.188
0.431
0.459
0.694
0.207
0.117
0.353
0.699
Partial Dependence
PAY_0
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

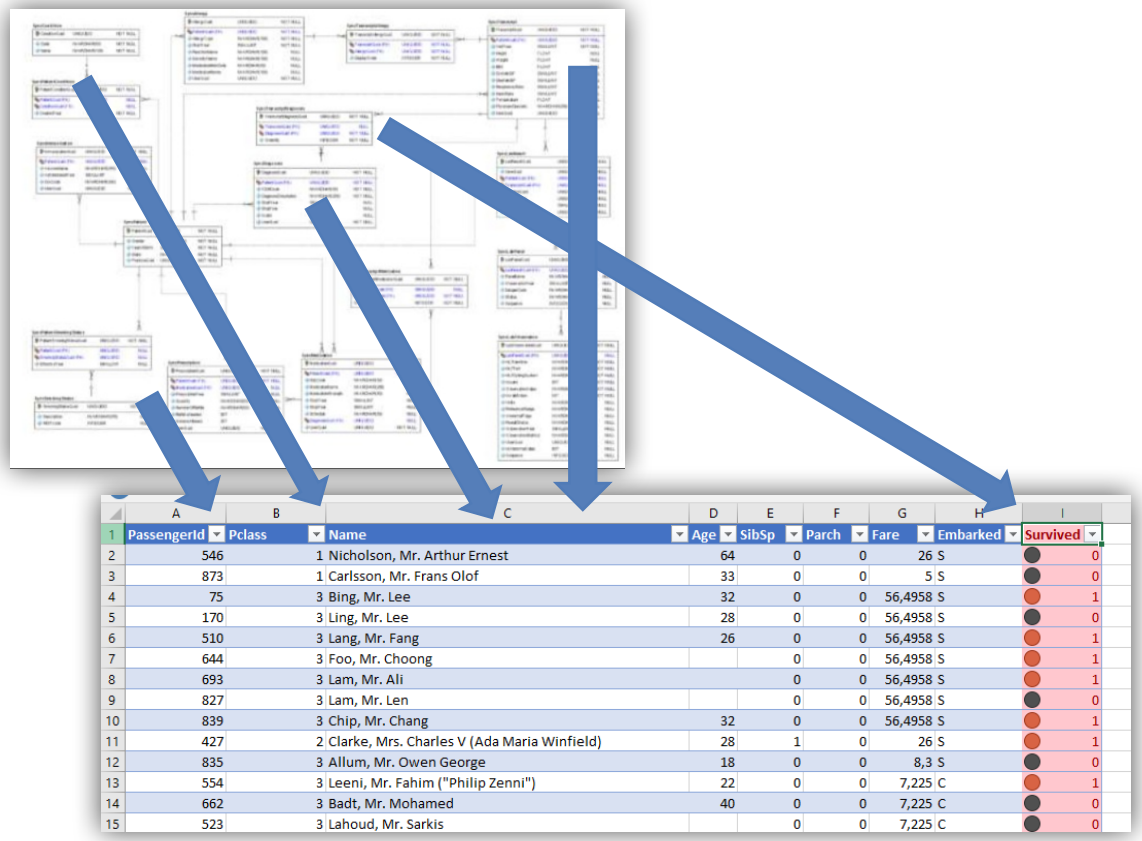
DataRobot

1. Ingest your data
2. Select the target variable
3. Build 100s of models in one click
4. Explore top models and get insights
5. Deploy best model and make predictions



What about Feature Discovery/Extraction?

- Machine Learning models only use a single table (with features & goal/label/target)



	A	B	C	D	E	F	G	H	I
1	PassengerId	Pclass	Name	Age	SibSp	Parch	Fare	Embarked	Survived
2	546	1	Nicholson, Mr. Arthur Ernest	64	0	0	26	S	0
3	873	1	Carlsson, Mr. Frans Olof	33	0	0	5	S	0
4	75	3	Bing, Mr. Lee	32	0	0	56,4958	S	1
5	170	3	Ling, Mr. Lee	28	0	0	56,4958	S	0
6	510	3	Lang, Mr. Fang	26	0	0	56,4958	S	1
7	644	3	Foo, Mr. Choong		0	0	56,4958	S	1
8	693	3	Lam, Mr. Ali		0	0	56,4958	S	1
9	827	3	Lam, Mr. Len		0	0	56,4958	S	0
10	839	3	Chip, Mr. Chang	32	0	0	56,4958	S	1
11	427	2	Clarke, Mrs. Charles V (Ada Maria Winfield)	28	1	0	26	S	1
12	835	3	Allum, Mr. Owen George	18	0	0	8,3	S	0
13	554	3	Leeni, Mr. Fahim ("Philip Zenni")	22	0	0	7,225	C	1
14	662	3	Badt, Mr. Mohamed	40	0	0	7,225	C	0
15	523	3	Lahoud, Mr. Sarkis		0	0	7,225	C	0

Harvard Business Review

DATA

Getting Value from Machine Learning Isn't About Fancier Algorithms — It's About Making It Easier to Use

by Ben Schreck, Max Kanter, Kalyan Veeramachaneni, Sanjeev Vohra, and Rajendra Prasad

MARCH 06, 2018

Feature Tools & Deep Feature Synthesis (MIT/DARPA)

Featuretools / **featuretools**

Watch 53 | Unstar 1,154 | Fork 110

"One of the holy grails of machine learning is to automate more and more of the feature engineering process." — Pedro Domingos, A Few Useful Things to Know about Machine Learning

```
Entityset: transactions
  Entities:
    customers (shape = [5, 3])
    sessions (shape = [35, 4])
    products (shape = [5, 2])
    transactions (shape = [500, 5])
  Relationships:
    transactions.product_id -> products.product_id
    transactions.session_id -> sessions.session_id
    sessions.customer_id -> customers.customer_id
```

Feature Tools & Deep Feature Synthesis (MIT/DARPA)

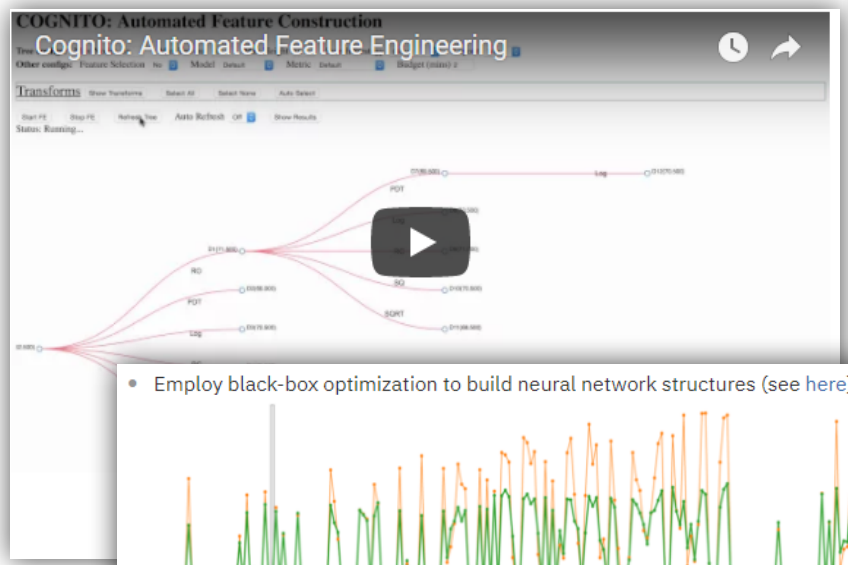
```
>> feature_matrix, features_defs = ft.dfs(entityset=es, target_entity="customers")
>> feature_matrix.head(5)
```

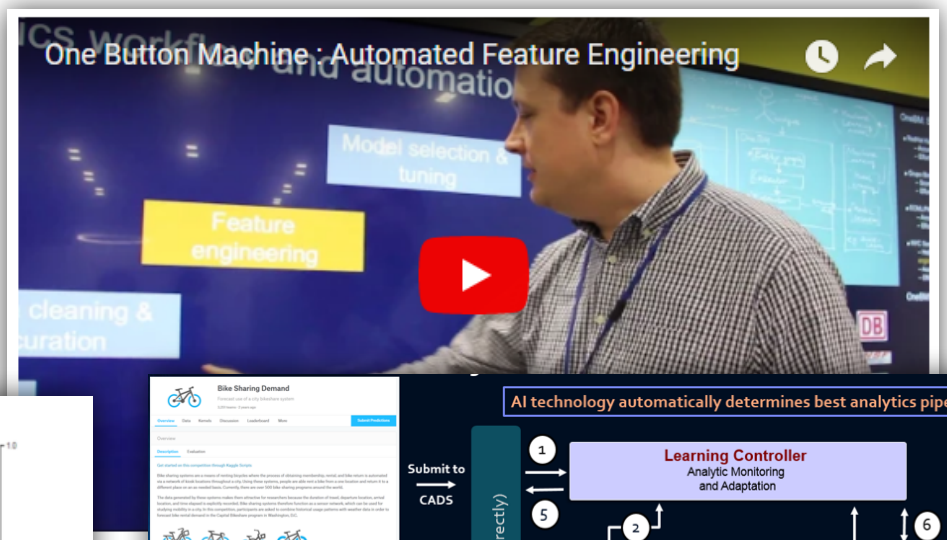
t))	STD(sessions.MAX(transactions.amount))	NUM_UNIQUE(sessions.DAY(session_start))	MIN(sessions.SKEW(transactions.am
150	5.857976	1	-0.
350	7.420480	1	-0.
976	12.537259	1	-0.
969	12.738488		
385	5.599228		

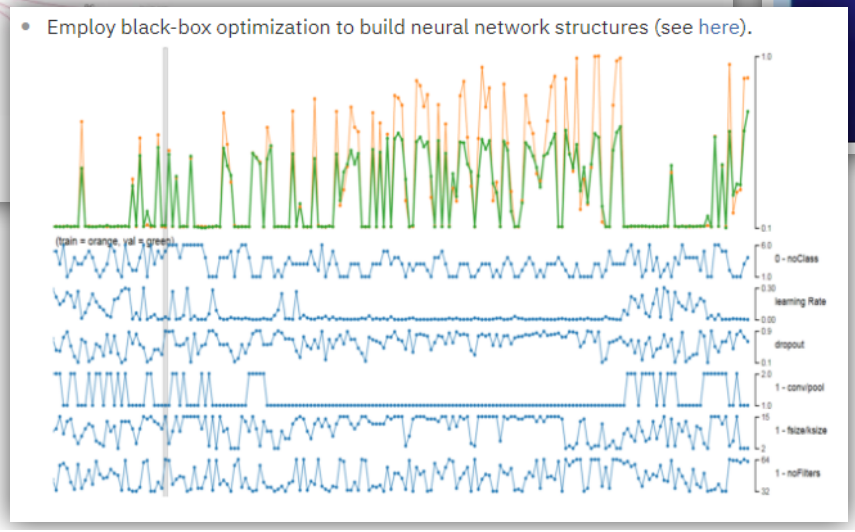
Cutoff times

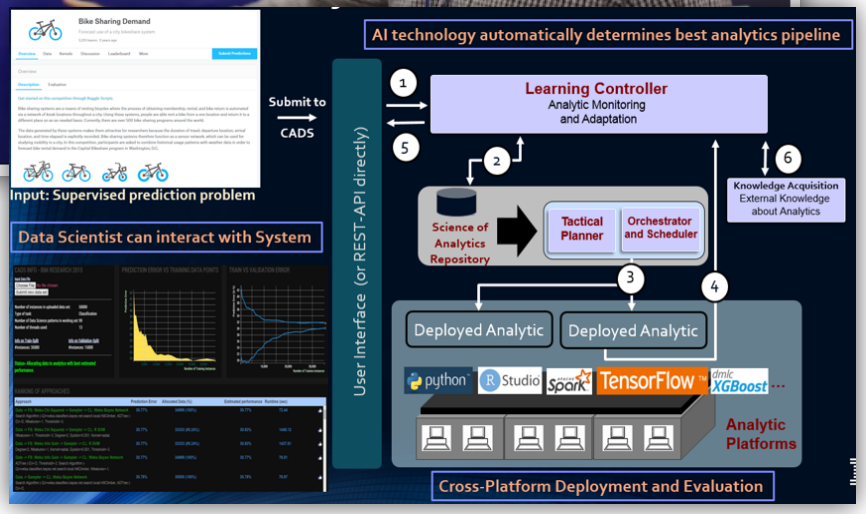
We can specify the time for each instance of the `target_entity` to calculate features. The timestamp represents the last time data can be used for calculating features. This is specified using a dataframe of cutoff times. Below we show an example of this dataframe for our customers example.

IBM R&D: One Button Machine, Cognito & more









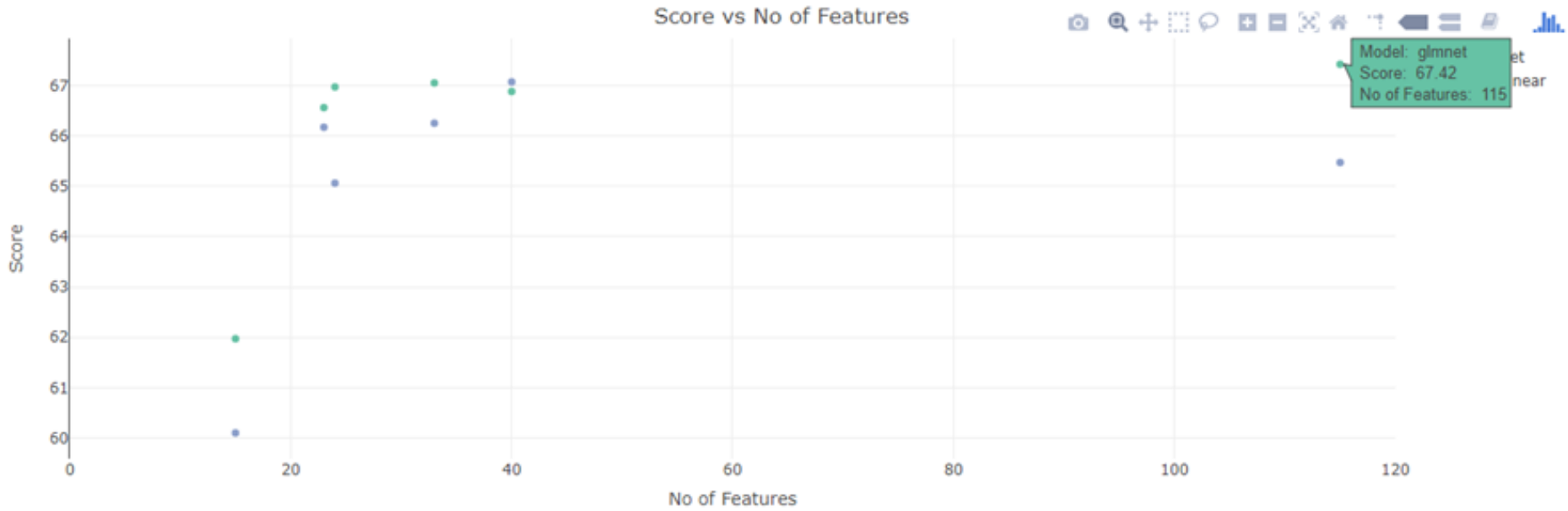
Project: Patient Conditions

Done
All jobs are finished

6/125/72
Tables/Features/Models

Azure/16/256
Cluster Mode/Nodes/Cores

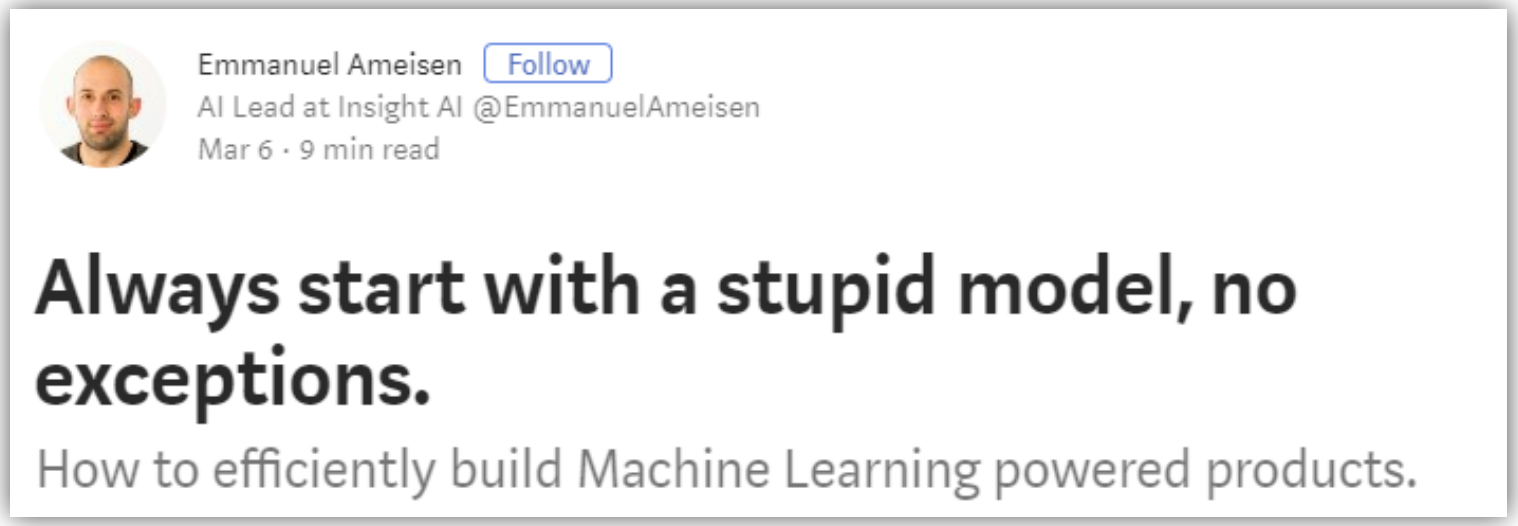
Jobs | Evaluation



AutoML Challenges

- “Overfitting” (+inf features), AUC/RMSE/… “hacking”
- Customization/Deployment/Commercial ”Lock-in”
- Excessive focus on best performance/Kaggle style
- ROI & Costs (commercial vs in-house)
- Misuse? (from non ML Experts)
- Complexity (open source tools)

Some Best Practices



Thank you!

Q&A

Should we use GUIs for Machine Learning?

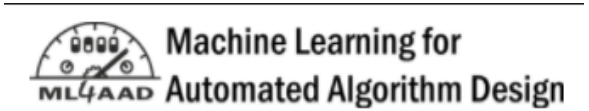
When to use Model/Hyper Params search tools ?

Will we be out of Jobs?

What cannot be automated? Where should we (humans) focus?

Thoughts? Let me know!

rui.quintino@devscope.net



Advisor

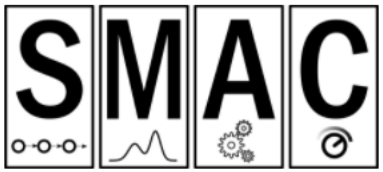
Hyperopt

Auto-WEKA

RoBO

automlk

mlrMBO



hyperband

ATM - Auto Tune Models