Θεσσαλονίκη, 21 Μαρτίου 2018

Digital Analytics

Meetup #6

Hello!

Τάσος Βεντούρης

Data Scientist and
Game Designer @
Hattrick Ltd

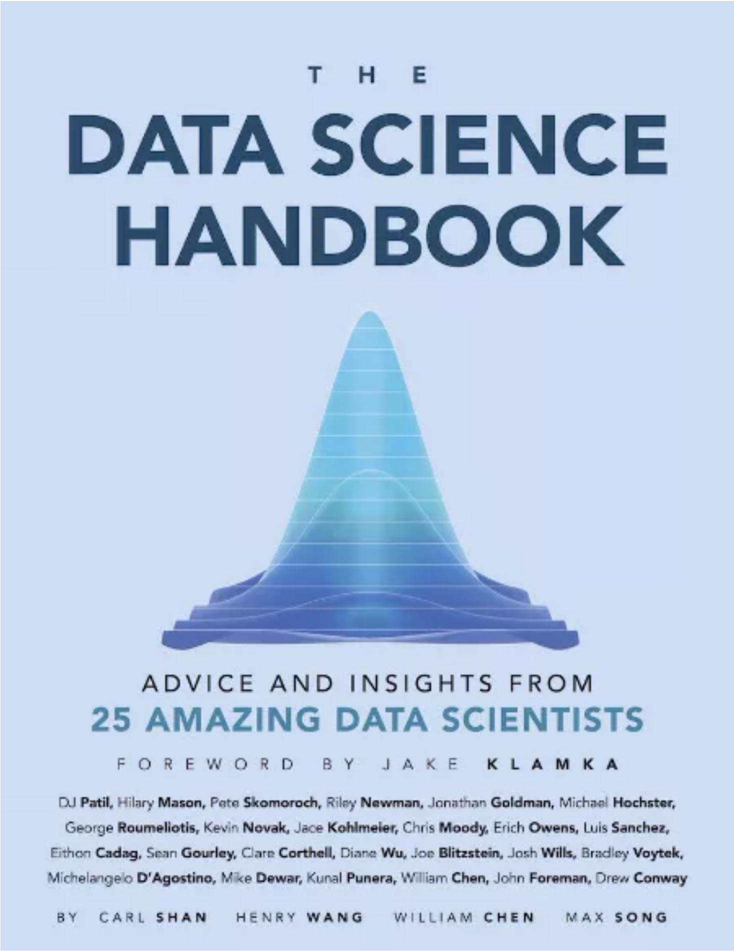
You can find me at:

@tasosventouris Tasos Ventouris

More About Me!

- (2012) BSc Mathematics @ Math Auth
- (2013) MSc Web Science @ Math Auth
- (2013) COO @ Open Knowledge Greece
- (2014) Mentor @ Open Knowledge Inter.
- (2014) Data Scientist @ Hattrick
- (2016) Found stackprime

The Data Science Handbook



Η ιστορία

Πως ξεκίνησαν όλα;

Timeline

- 1960 - Computer Science = Data Science από Peter Naur
- 1974 - Πρώτη φορά σε δημοσίευση από Peter Naur
- 1996 - Συνέδριο με τίτλο “Data Science, classification, and related methods”
- 1997 - Ομιλία του Jeff Wu με τίτλο “Statistics = Data Science?”
- 2001 - William S. Cleveland χρησιμοποίησε τη Data Science ως ανεξάρτητο όρο σε άρθρο της “International Statistical Review”
- 2002 - Committee on Data for Science & Technology. Νέο περιοδικό με τίτλο Data Science Journal
- 2003 - The Journal of Data Science από Columbia University
- 2008 - DJ Patil & Jeff Hammerbacher χρησιμοποίησαν τον τίτλο Data Scientist
- 2012 - Άρθρο από Harvard Business Review με τίτλο “Data Scientist: The Sexiest Job of the 21st Century”

Data Science

Bubble or not?

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

MENU

Harvard Business Review

DATA

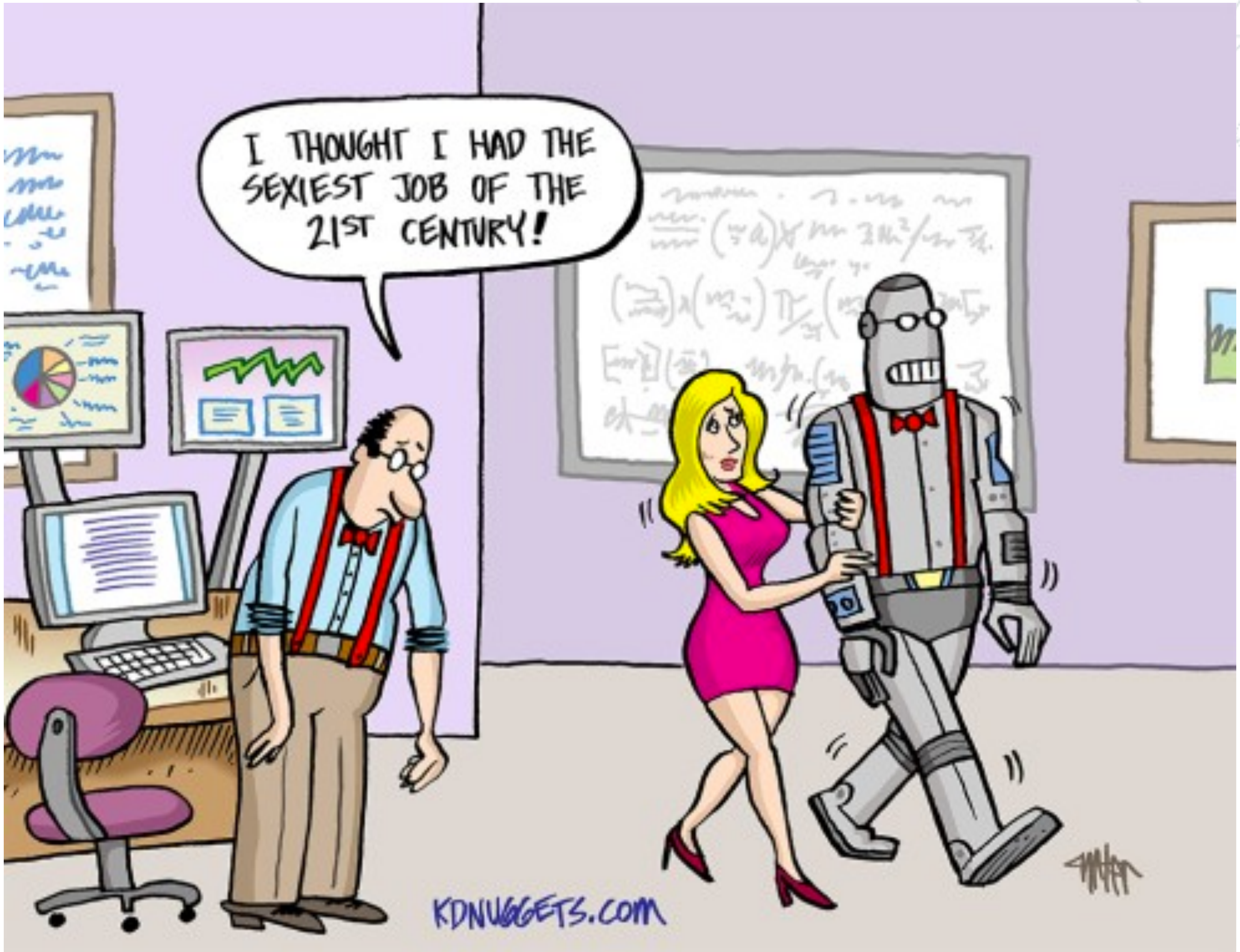
Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY | SAVE | SHARE | 4 COMMENT | TEXT SIZE | PRINT | $8.95 BUY COPIES

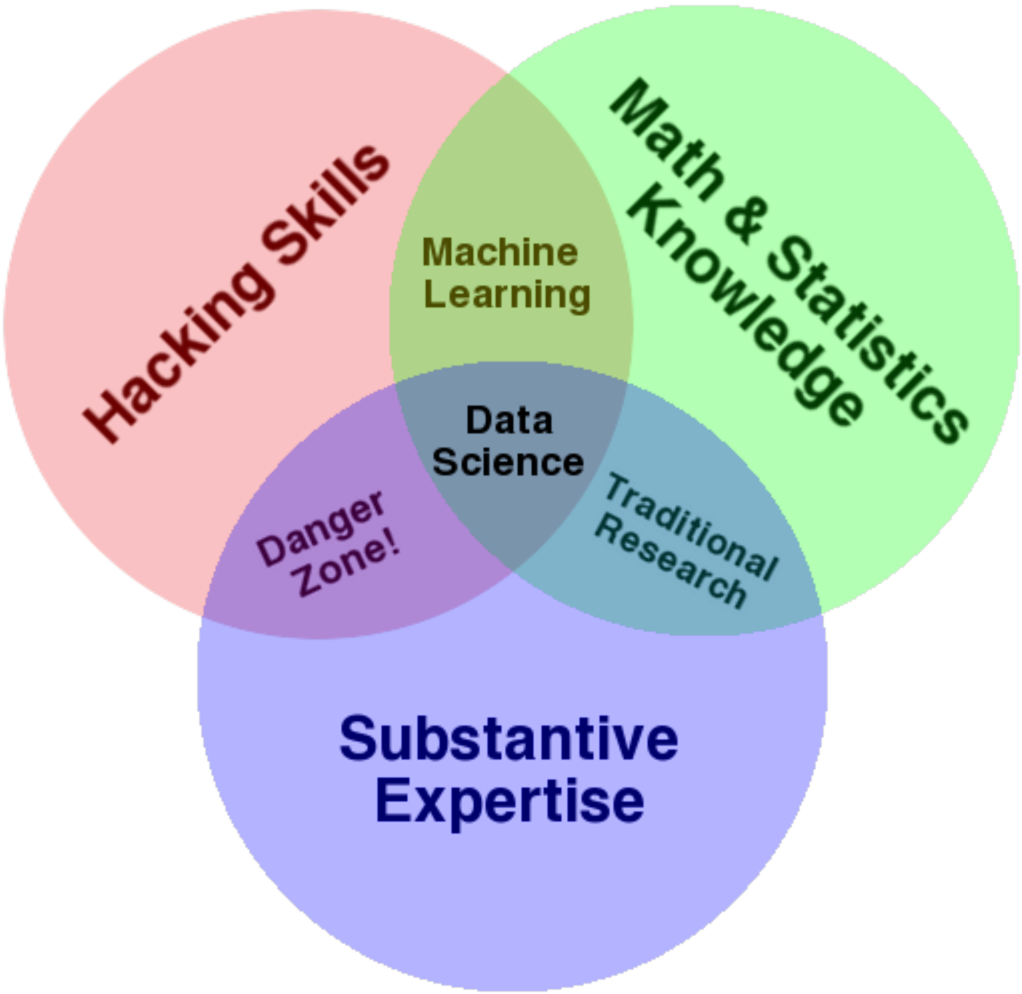
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join.


I THOUGHT I HAD THE SEXIEST JOB OF THE 21ST CENTURY!
KDNUGGETS.COM

“

The creation of data products

*Data product = Ένα εργαλείο που δημιουργήθηκε με τη χρήση δεδομένων και βοηθάει στη λήψη αποφάσεων.


Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

Data Scientist

Ποια είναι τα χαρακτηριστικά του;

Josh Wills
@josh_wills

Follow

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

“

A Data Scientist is a ***statistician*** *who lives in San Francisco* 😋

A Data Scientist is a person who is able to…

A Data Scientist is a person who is able to…

run a regression

A Data Scientist is a person who is able to…

run a regression
write a sql query

A Data Scientist is a person who is able to…

run a regression
write a sql query
scrape a web site

A Data Scientist is a person who is able to…

run a regression
write a sql query
scrape a web site
design an experiment

A Data Scientist is a person who is able to…

run a regression
write a sql query
scrape a web site
design an experiment
factor matrices

A Data Scientist is a person who is able to…

run a regression
write a sql query
scrape a web site
design an experiment
factor matrices
use a data frame

A Data Scientist is a person who is able to…

run a regression
write a sql query
scrape a web site
design an experiment
factor matrices
use a data frame
pretend to understand
deep learning

A Data Scientist is a person who is able to…

run a regression
write a sql query
scrape a web site
design an experiment
factor matrices
use a data frame
pretend to understand
deep learning
steal from the d3 gallery

A Data Scientist is a person who is able to…

run a regression
write a sql query
scrape a web site
design an experiment
factor matrices
use a data frame
pretend to understand deep learning
steal from the d3 gallery

A Data Scientist is a person who is able to…

run a regression
write a sql query
scrape a web site
design an experiment
factor matrices
use a data frame
pretend to understand deep learning
steal from the d3 gallery
argue r versus python

A Data Scientist is a person who is able to…

run a regression
write a sql query
scrape a web site
design an experiment
factor matrices
use a data frame
pretend to understand
deep learning
steal from the d3 gallery
argue r versus python
use mapreduce

A Data Scientist is a person who is able to…

run a regression
write a sql query
scrape a web site
design an experiment
factor matrices
use a data frame
pretend to understand deep learning
steal from the d3 gallery
argue r versus python
use mapreduce
build a dashboard

A Data Scientist is a person who is able to…

run a regression
write a sql query
scrape a web site
design an experiment
factor matrices
use a data frame
pretend to understand deep learning
steal from the d3 gallery
argue r versus python
use mapreduce
build a dashboard
clean up messy data

A Data Scientist is a person who is able to…

run a regression
write a sql query
scrape a web site
design an experiment
factor matrices
use a data frame
pretend to understand deep learning
steal from the d3 gallery
argue r versus python
use mapreduce
build a dashboard
clean up messy data
test a hypothesis

A Data Scientist is a person who is able to…

run a regression
write a sql query
scrape a web site
design an experiment
factor matrices
use a data frame
pretend to understand
deep learning
steal from the d3 gallery

argue r versus python
use mapreduce
build a dashboard
clean up messy data
test a hypothesis
script a shell

A Data Scientist is a person who is able to…

run a regression
write a sql query
scrape a web site
design an experiment
factor matrices
use a data frame
pretend to understand
deep learning
steal from the d3 gallery

argue r versus python
use mapreduce
build a dashboard
clean up messy data
test a hypothesis
script a shell
hack a p-value

A Data Scientist is a person who is able to…

run a regression
write a sql query
scrape a web site
design an experiment
factor matrices
use a data frame
pretend to understand
deep learning
steal from the d3 gallery
argue r versus python
use mapreduce
build a dashboard
clean up messy data
test a hypothesis
script a shell
hack a p-value
machine-learn a model

A Data Scientist is a person who is able to…

run a regression
write a sql query
scrape a web site
design an experiment
factor matrices
use a data frame
pretend to understand
deep learning
steal from the d3 gallery

argue r versus python
use mapreduce
build a dashboard
clean up messy data
test a hypothesis
script a shell
hack a p-value
machine-learn a model
talk to a business person

A Data Scientist is a person who is able to…

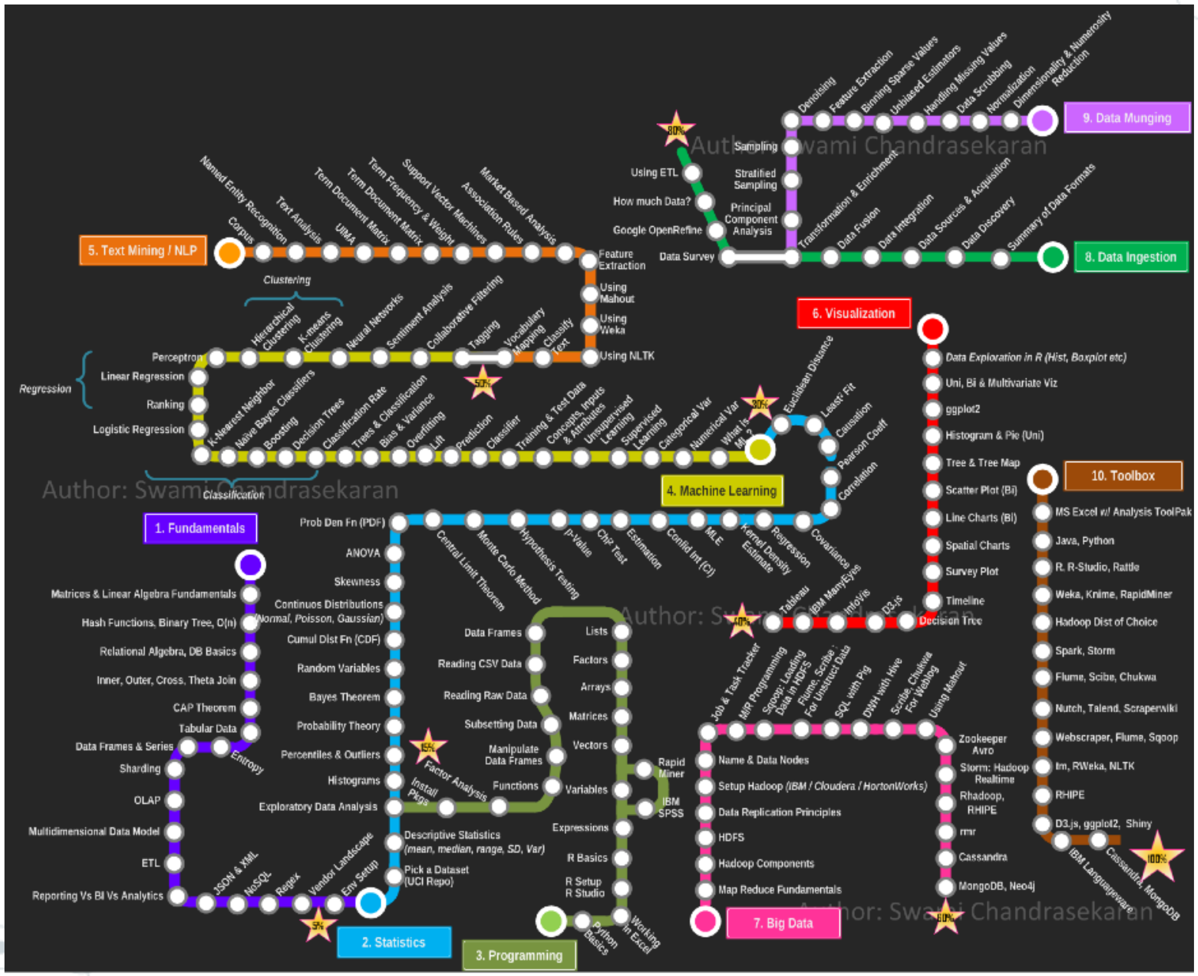
run a regression
write a sql query
scrape a web site
design an experiment
factor matrices
use a data frame
pretend to understand
deep learning
steal from the d3 gallery

argue r versus python
use mapreduce
build a dashboard
clean up messy data
test a hypothesis
script a shell
hack a p-value
machine-learn a model
talk to a business person

Το μονοπάτι ενός Data Scientist

RPG character?

Το μονοπάτι του Data Scientist

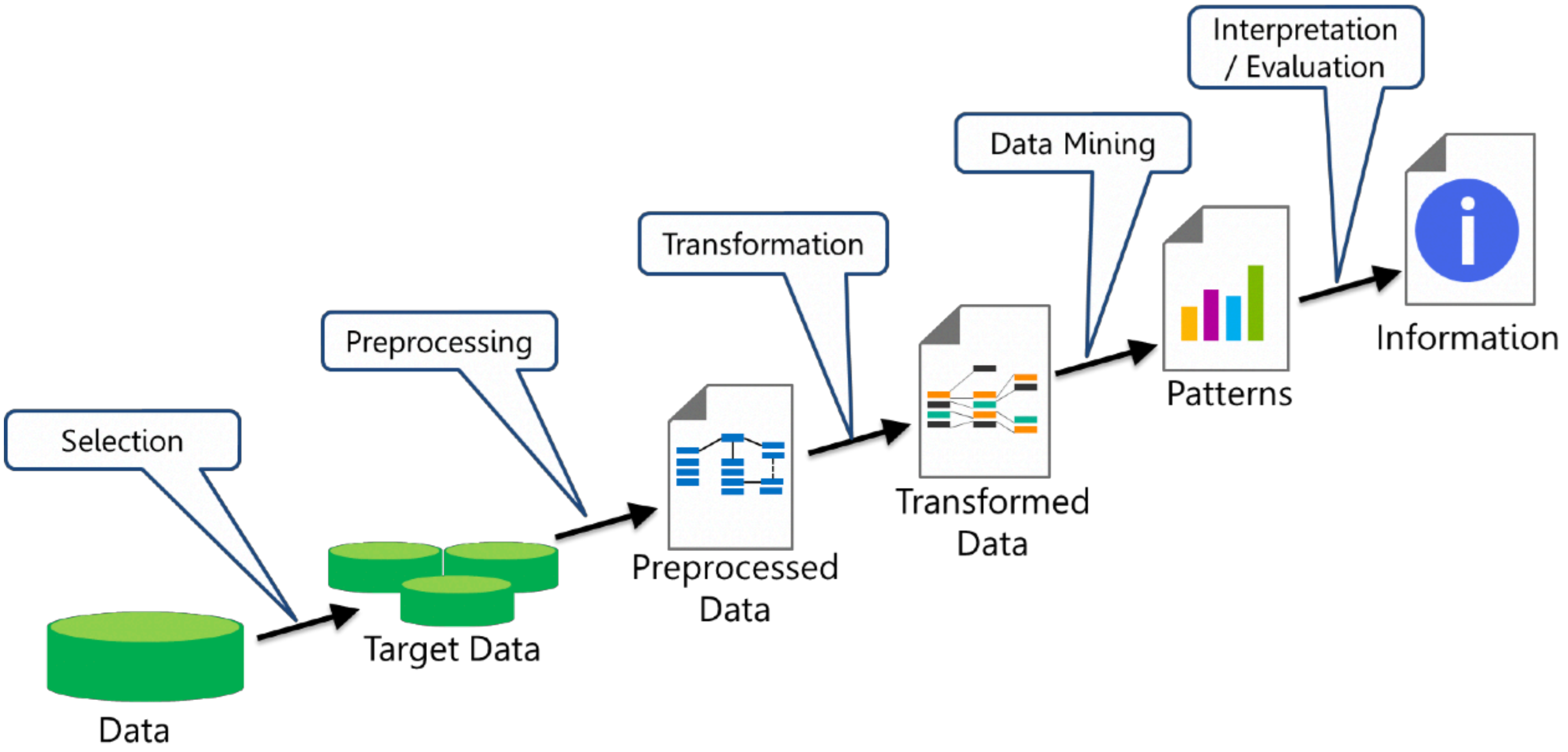


Use your brain to take decisions
Don’t use it to store info

Data Science Process

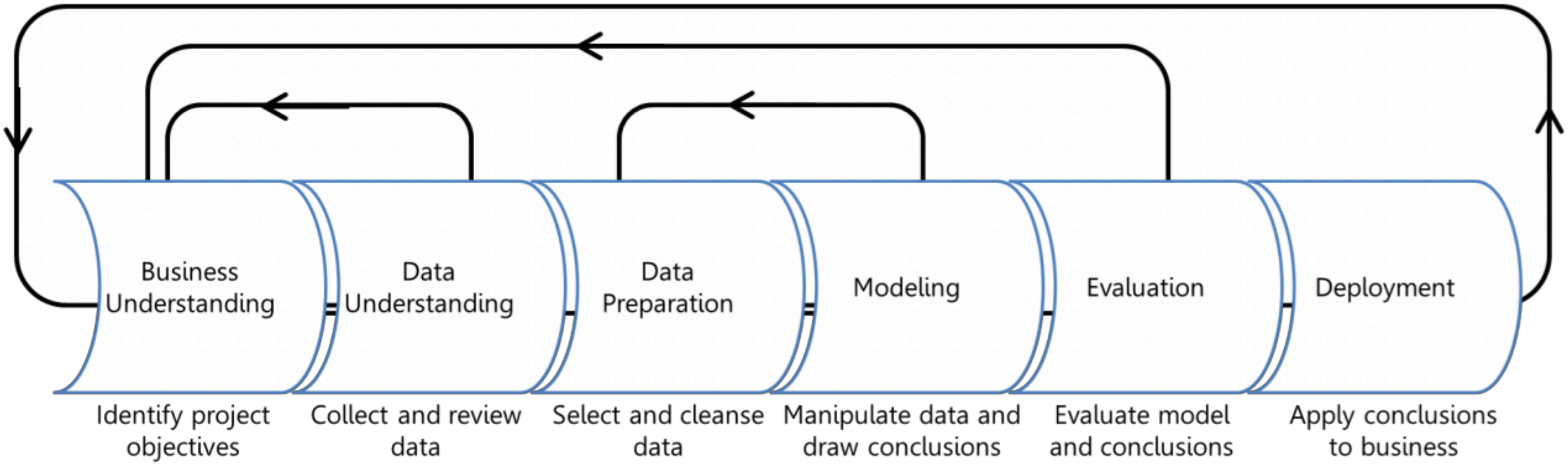
Ποια είναι τα βήματα;

To 1997



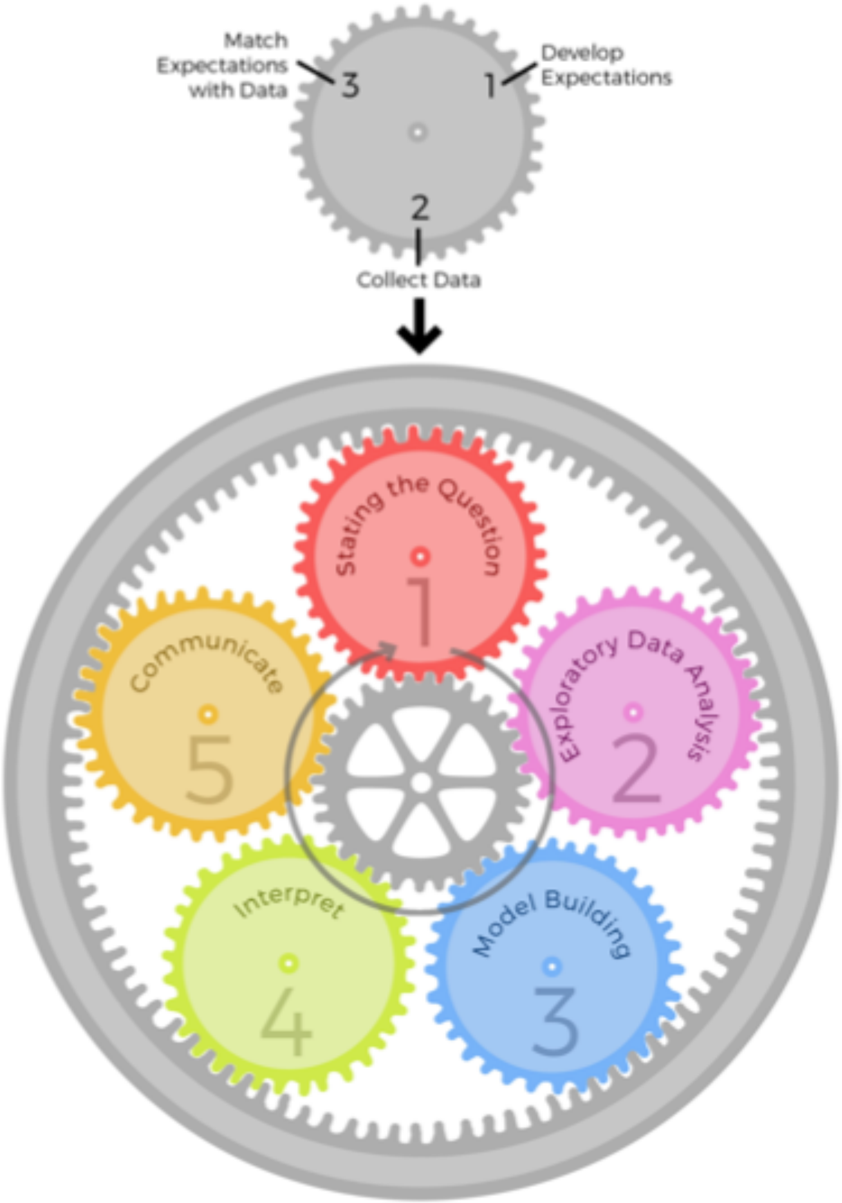
Selection
Preprocessing
Transformation
Data Mining
Interpretation / Evaluation
Data
Target Data
Preprocessed Data
Transformed Data
Patterns
Information

To 2000



Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Identify project objectives
Collect and review data
Select and cleanse data
Manipulate data and draw conclusions
Evaluate model and conclusions
Apply conclusions to business

Reality???

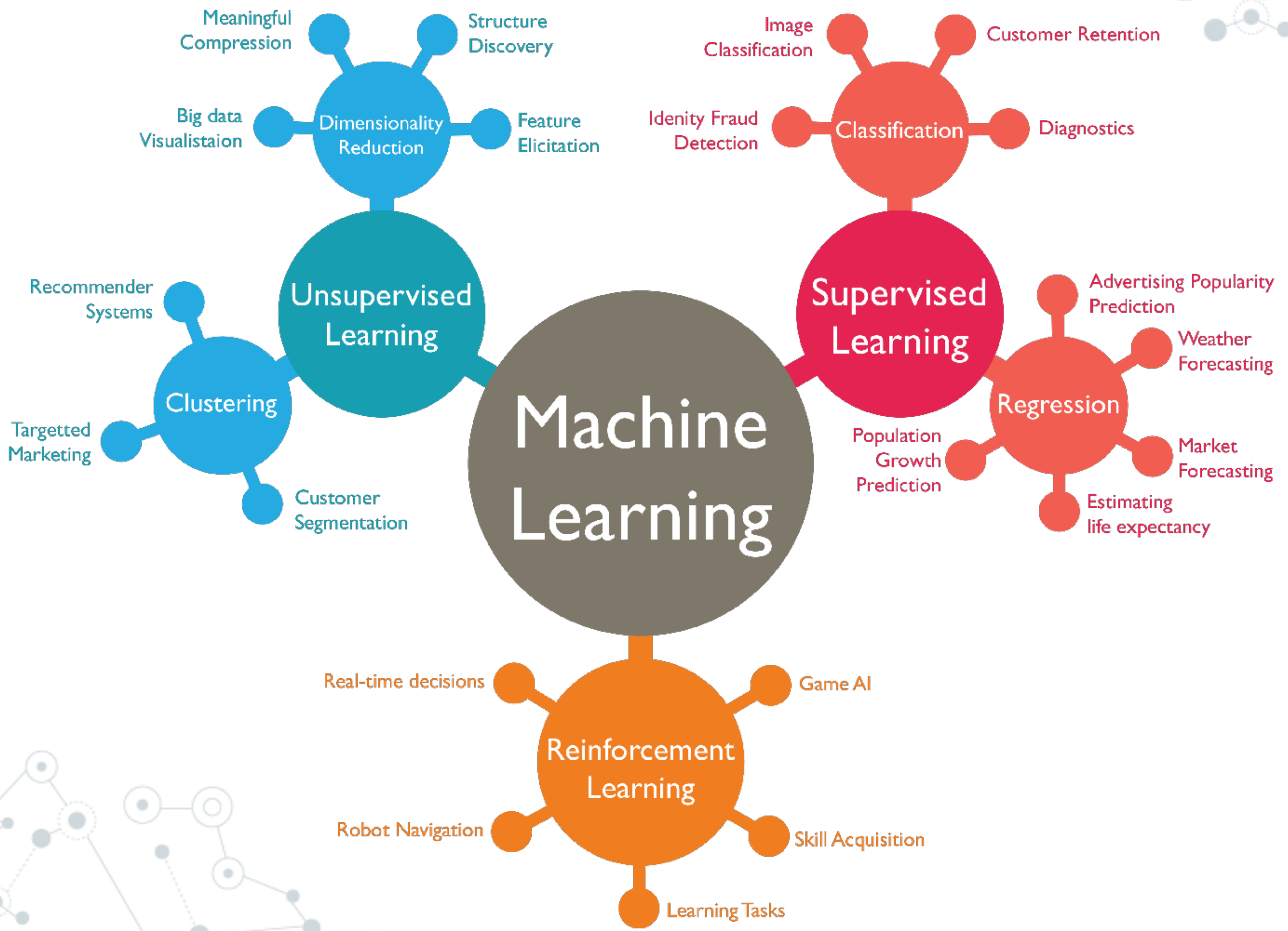


Match
Expectations
with Data
3
1
Develop
Expectations
2
Collect Data
Stating the Question
1
Exploratory Data Analysis
2
Model Building
3
Interpret
4
Communicate
5

Machine Learning

Ή αλλιώς μηχανές μάθησης

Machine Learning family

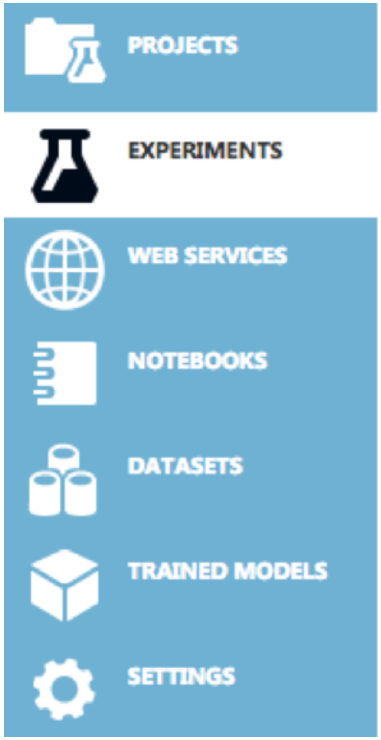


Azure ML

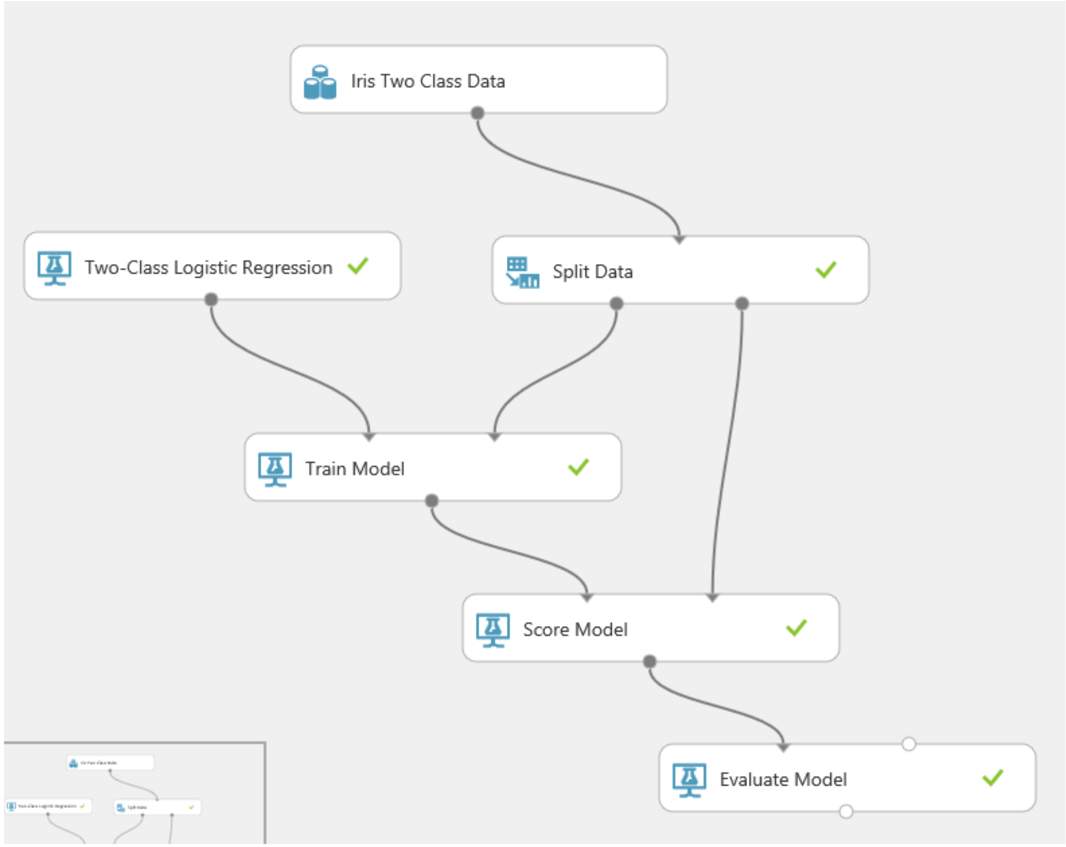
https://studio.azureml.net

What is Azure Machine Learning?

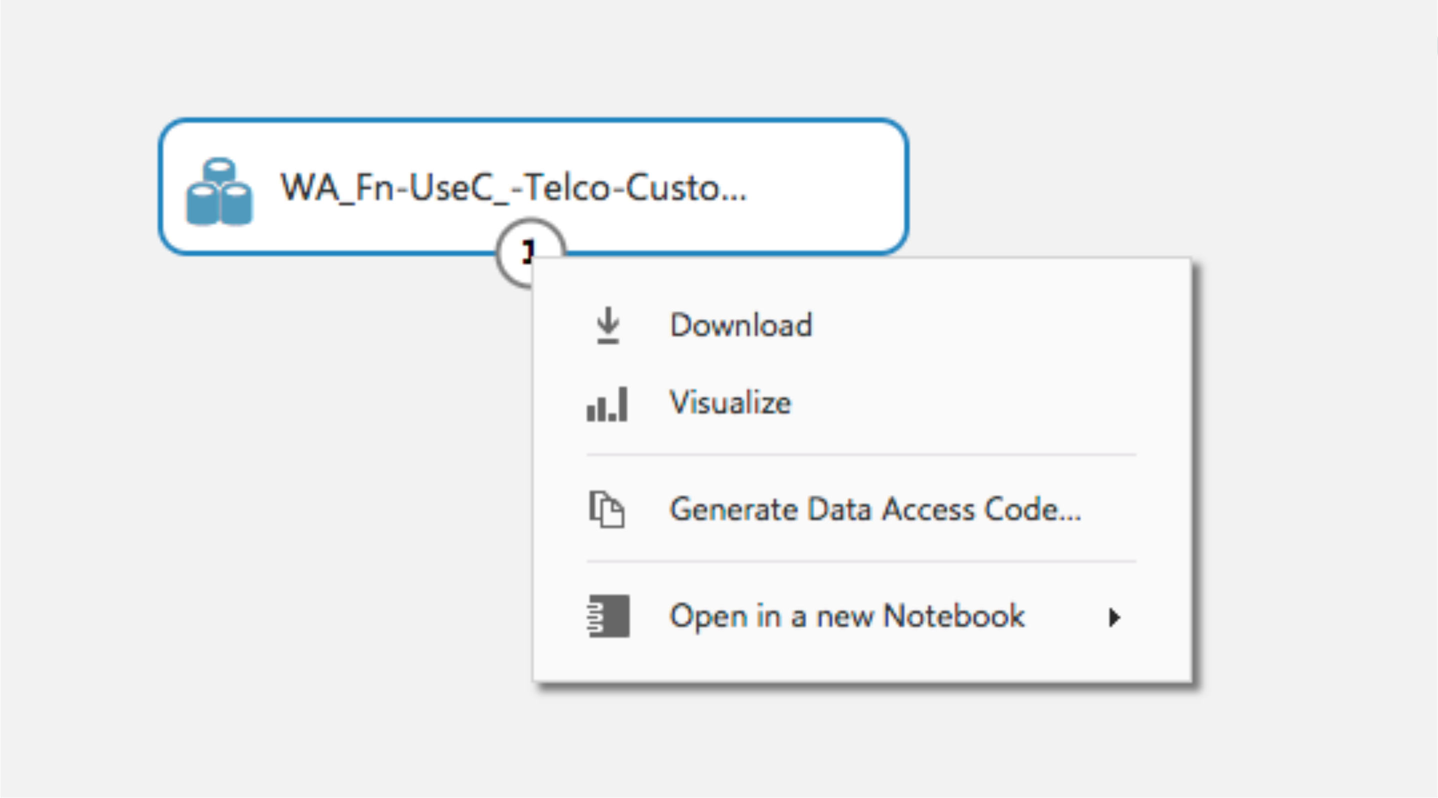
- One of the best tools to learn ML
- Drag and drop
- R, Python, SQL, Jupyter integration
- No code (if you don't want it)
- Web Service (RESTful API)
- Part of Azure Cloud



Azure ML Modules



User interaction



Dataset Preview

Churn Rate Telco › WA_Fn-UseC_-Telco-Customer-Churn.csv › dataset

rows	columns
7043	21

view as

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetSe
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL
7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL
9237-HQITU	Female	0	No	No	2	Yes	No	Fiber opti
9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber opti
1452-	Male	0	No	Yes	22	Yes	Yes	Fiber opti

Dataset Preview

Churn Rate Telco › WA_Fn-UseC_-Telco-Customer-Churn.csv › dataset

rows	columns
7043	21

view as

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetSe
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL
7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL
9237-HQITU	Female	0	No	No	2	Yes	No	Fiber opti
9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber opti
1452-	Male	0	No	Yes	22	Yes	Yes	Fiber opti

Dataset Preview

Churn Rate Telco › WA_Fn-UseC_-Telco-Customer-Churn.csv › dataset

rows	columns
7043	21

view as

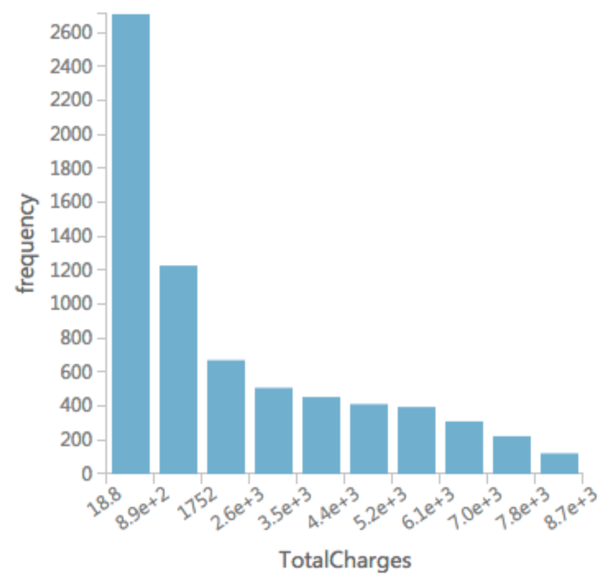
customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetSe
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL
7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL
9237-HQITU	Female	0	No	No	2	Yes	No	Fiber opti
9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber opti
1452-	Male	0	No	Yes	22	Yes	Yes	Fiber opti

Dataset Preview

PaymentMethod	MonthlyCharges	TotalCharges	Ch
Electronic check	29.85	29.85	Nc
Mailed check	56.95	1889.5	Nc
Mailed check	53.85	108.15	Ye
Bank transfer (automatic)	42.3	1840.75	Nc
Electronic check	70.7	151.65	Ye
Electronic check	99.65	820.5	Ye
Credit card	89.1	1949.4	Nc

Visualizations

TotalCharges
Histogram



Dataset Preview

PaymentMethod	MonthlyCharges	TotalCharges	Churn
Electronic check	29.85	29.85	No
Mailed check	56.95	1889.5	No
Mailed check	53.85	108.15	Yes
Bank transfer (automatic)	42.3	1840.75	No
Electronic check	70.7	151.65	Yes

Statistics

Mean	2283.3004
Median	1397.475
Min	18.8
Max	8684.8
Standard Deviation	2266.7714
Unique Values	6530
Missing Values	11
Feature Type	Numeric Feature

Visualizations

Dataset Preview

PaymentMethod	MonthlyCharges	TotalCharges	Churn
Electronic check	29.85	29.85	No
Mailed check	56.95	1889.5	No
Mailed check	53.85	108.15	Yes
Bank transfer (automatic)	42.3	1840.75	No
Electronic check	70.7	151.65	Yes

Statistics

Mean	2283.3004
Median	1397.475
Min	18.8
Max	8684.8
Standard Deviation	2266.7714
Unique Values	6530
Missing Values	11
Feature Type	Numeric Feature

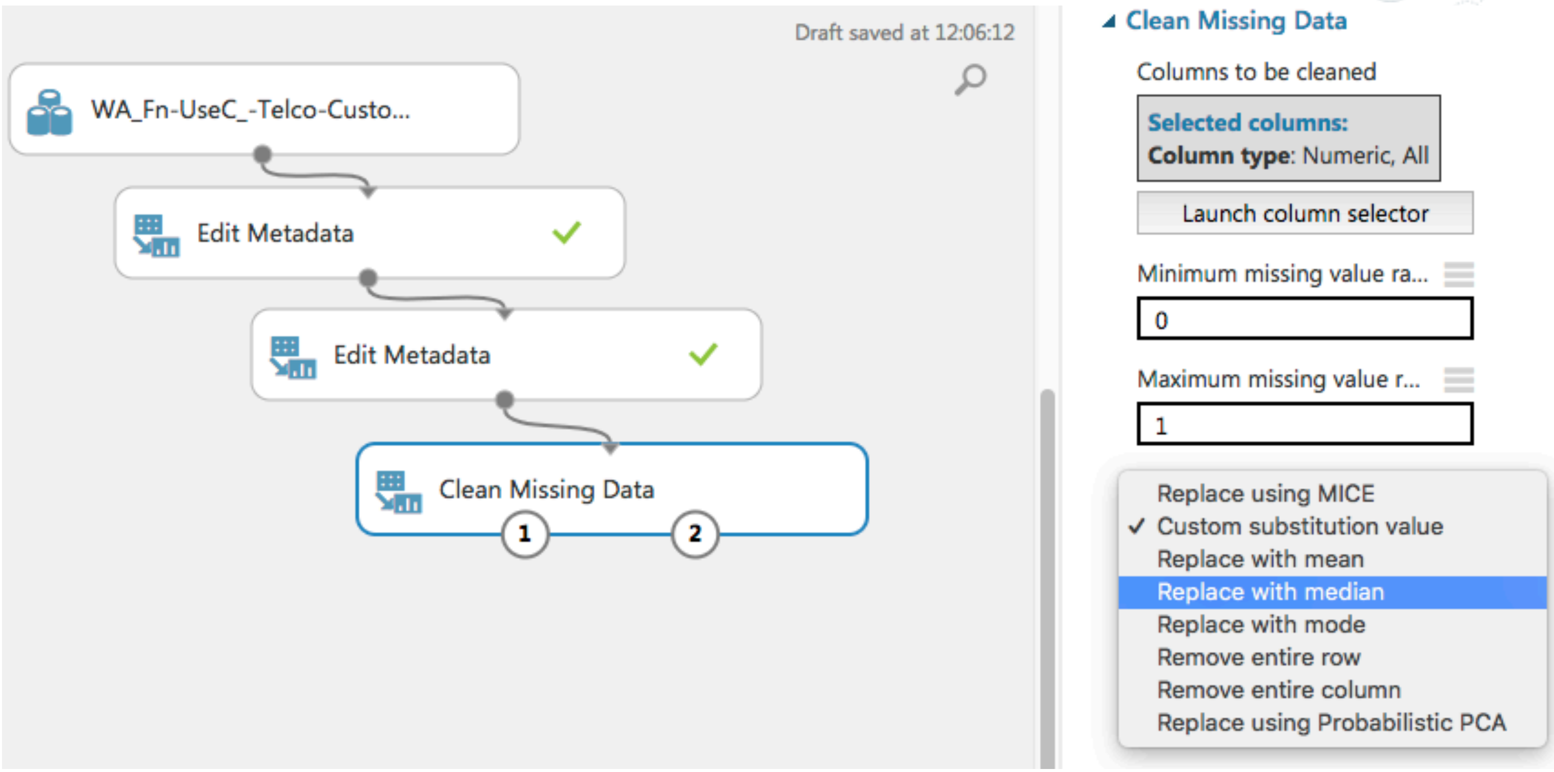
Visualizations

Data Transformation

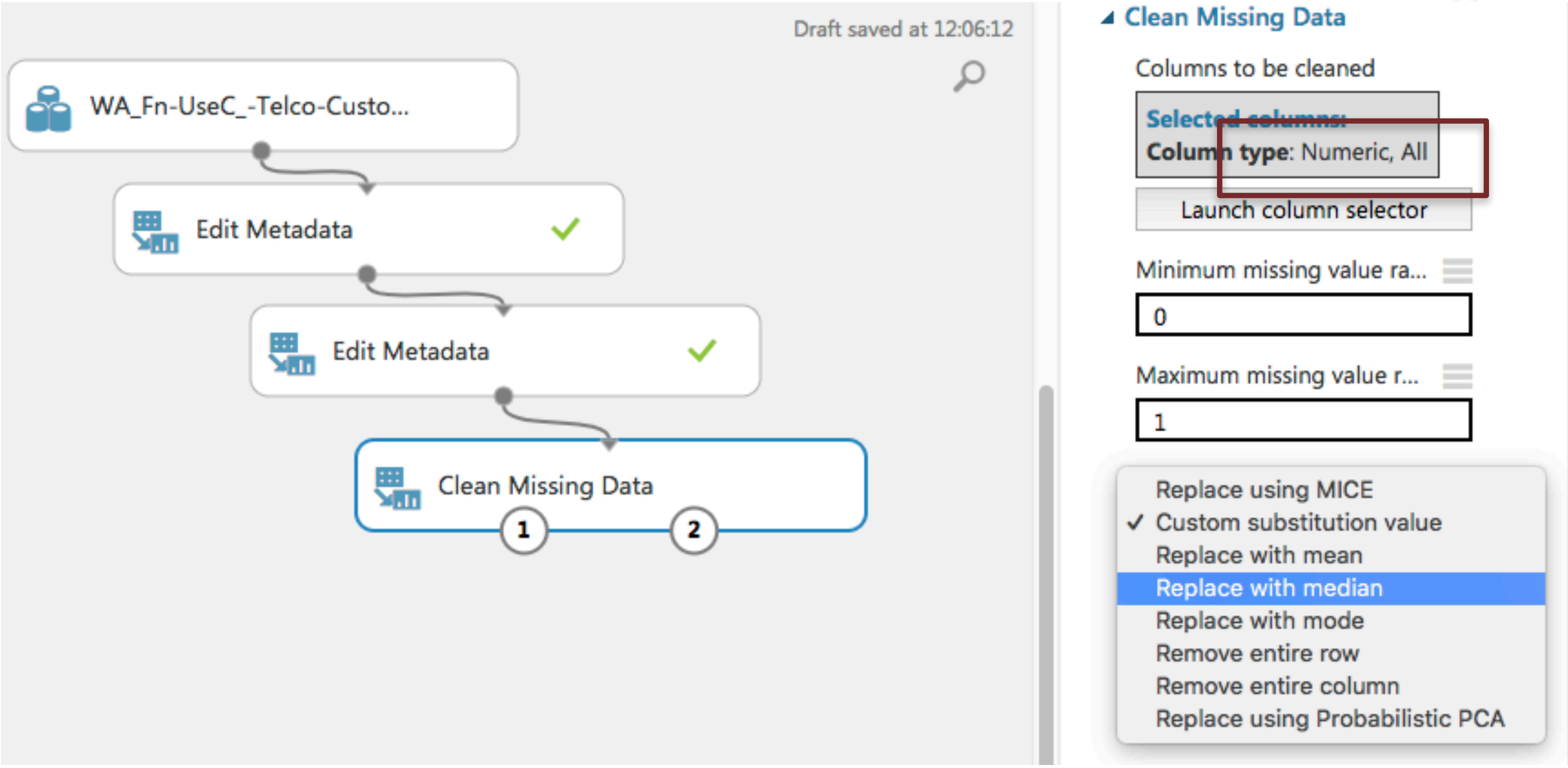
Data Transformation

- Filters (like median and moving average)
- Manipulation (add/remove/edit/join data)
- Sample/Partition/Split
- Scale and Reduce (Normalize, PCA)

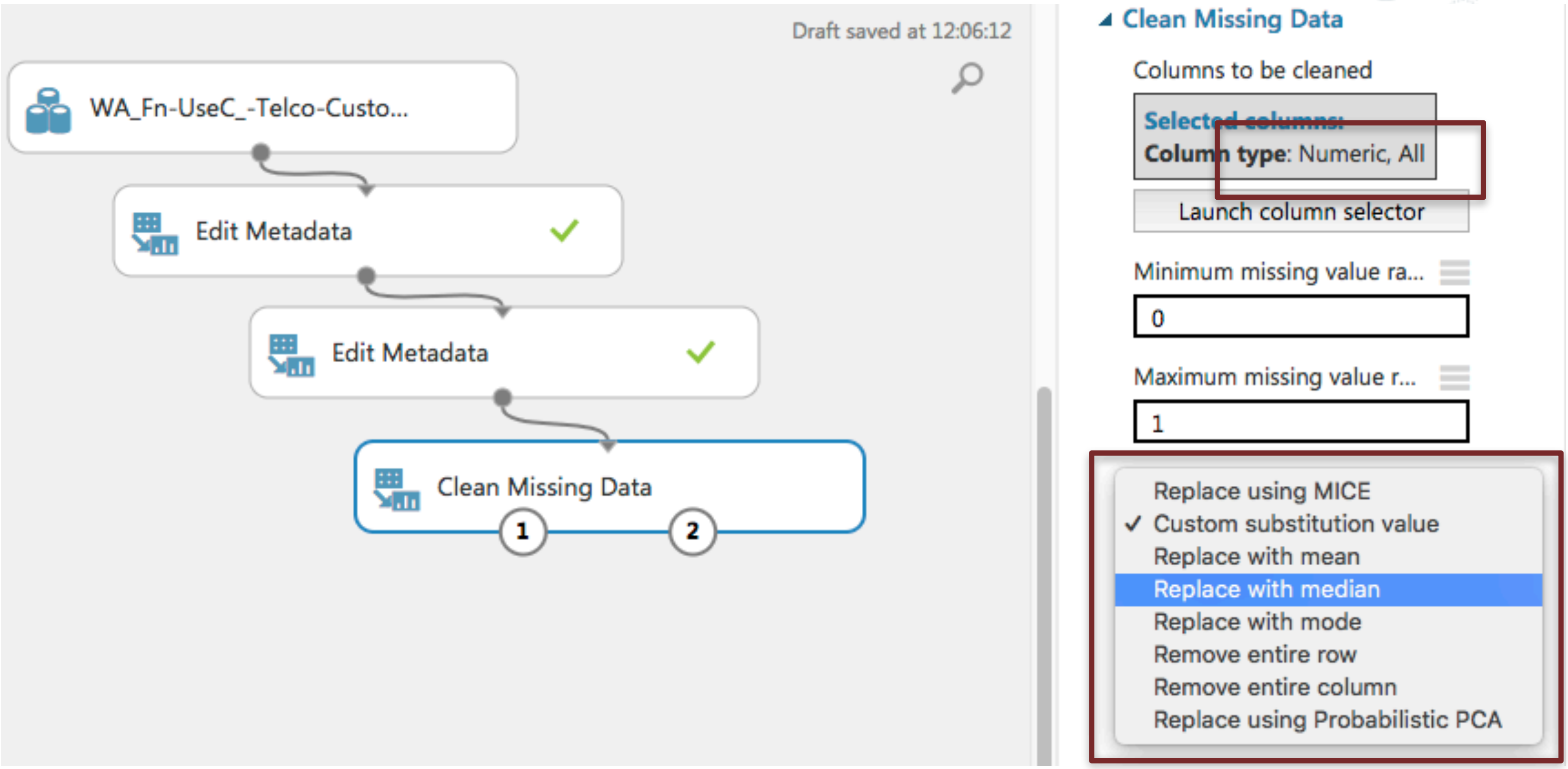
Missing Data



Missing Data



Missing Data



Who said anything about PCA?

Properties Project

Principal Component Analysis

Selected columns

Selected columns:
All columns
Exclude column names: Churn

Launch column selector

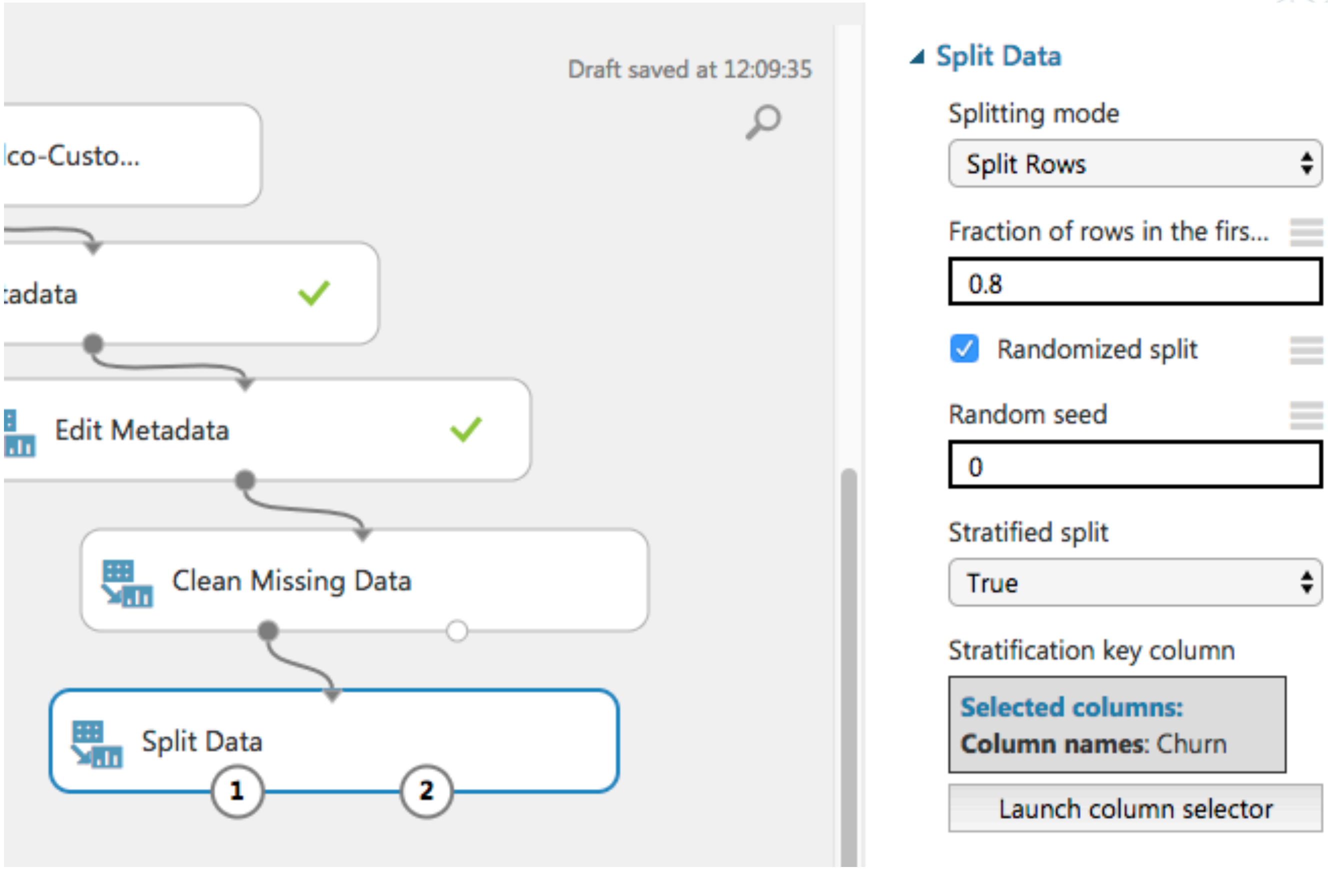
Number of dimensions to reduce to

5

☑ Normalize dense dataset to zero mean

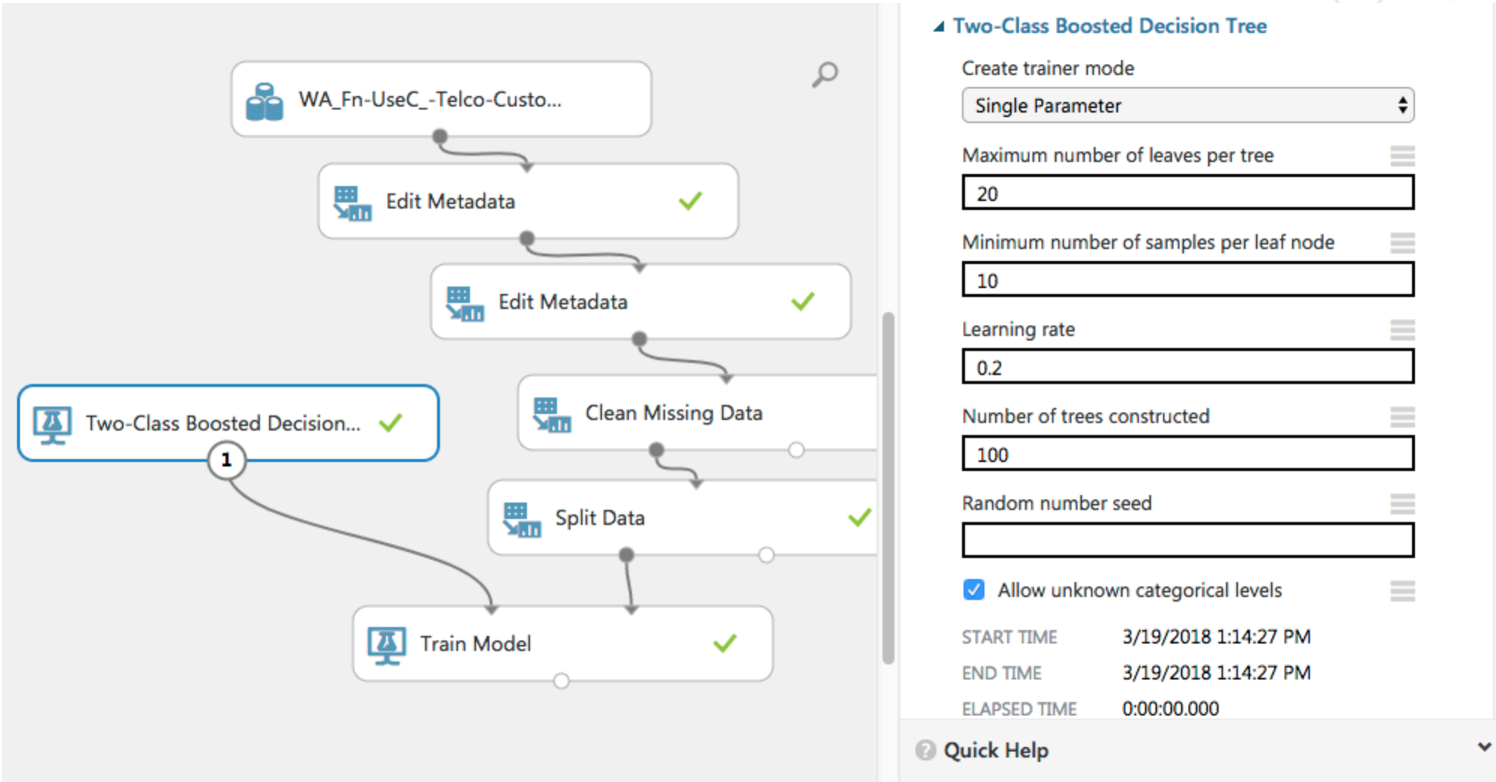
Churn	Col1	Col2	Col3	Col4	Col5
No	-30.114524	-26.252327	-13.708871	-1.2272	0.131734
No	-1890.149954	-28.87522	-33.194957	-0.027303	-0.109888
Yes	-108.624725	-47.016248	-24.223736	0.007596	0.026025
No	-1841.375793	-11.064167	-37.19584	-1.000805	-0.206422
Yes	-152.266892	-61.998133	-31.166208	0.691678	0.834274
Yes	-821.369645	-82.763106	-41.403142	2.23297	0.407882
No	-1950.216186	-62.811068	-35.933343	1.388941	0.096455
No	-302.22283	-21.135979	-18.480932	-1.283869	0.127006
Yes	-3046.970944	-70.070366	-35.254616	2.144601	-0.193108
No	-3488.719888	-9.714618	-39.038345	-0.455416	-0.216679
No	-587.952346	-36.834956	-26.713525	-0.171909	-1.112032
No	-327.081096	-8.594353	-18.7961	2.372951	0.482018
No	-5682.046789	-42.791289	-28.989216	1.617872	0.225253
Yes	-5037.245107	-52.244455	-30.015284	1.695195	1.1711
No	-2686.977808	-73.381479	-37.10689	2.192411	-0.0909
No	-7896.135576	-41.037799	-18.313305	2.017786	-1.204476
No	-1023.513935	8.407268	-43.570055	2.160647	0.910689
No	-7383.236219	-36.257137	-23.291982	1.593812	0.046805
Yes	-528.875633	-43.075567	-27.043751	0.347563	-1.821551
No	-1863.722043	-64.440662	-36.463617	1.553473	0.647516

Split Data

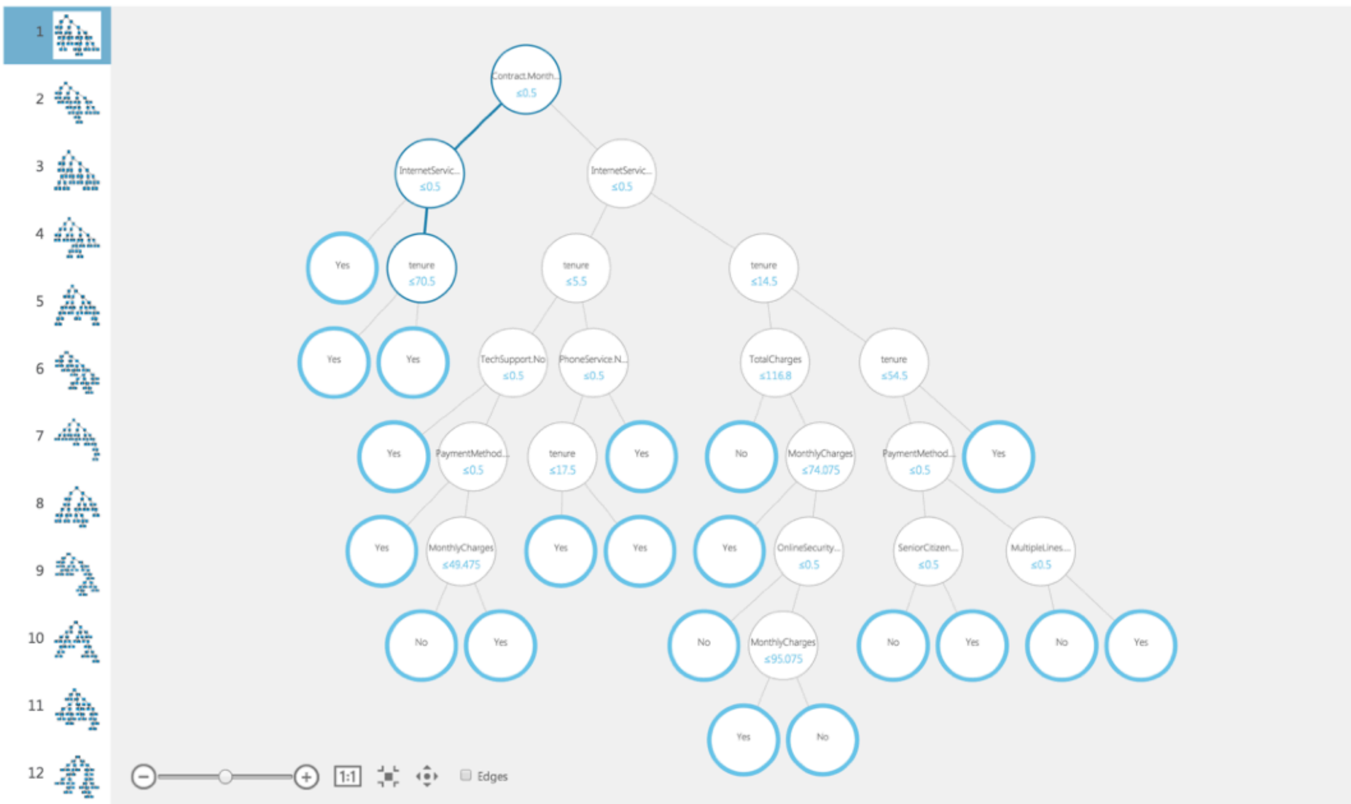


Draft saved at 12:09:35
lco-Custo...
tadata
Edit Metadata
Clean Missing Data
Split Data
1
2
Split Data
Splitting mode
Split Rows
Fraction of rows in the firs...
0.8
Randomized split
Random seed
0
Stratified split
True
Stratification key column
Selected columns:
Column names: Churn
Launch column selector

Train Model


WA_Fn-UseC_-Telco-Custo...
Edit Metadata
Edit Metadata
Two-Class Boosted Decision...
1
Clean Missing Data
Split Data
Train Model
Two-Class Boosted Decision Tree
Create trainer mode
Single Parameter
Maximum number of leaves per tree
20
Minimum number of samples per leaf node
10
Learning rate
0.2
Number of trees constructed
100
Random number seed
Allow unknown categorical levels
START TIME 3/19/2018 1:14:27 PM
END TIME 3/19/2018 1:14:27 PM
ELAPSED TIME 0:00:00.000
Quick Help

Visualise the results



Statistics

Contract.Month-to-month

PREDICATE Contract.Month-to-month ≤ 0.5

SPLIT GAIN 28.876266

InternetService.Fiber optic

PREDICATE InternetService.Fiber optic ≤ 0.5

SPLIT GAIN 0.909322

GAIN P-VALUE 0.000002

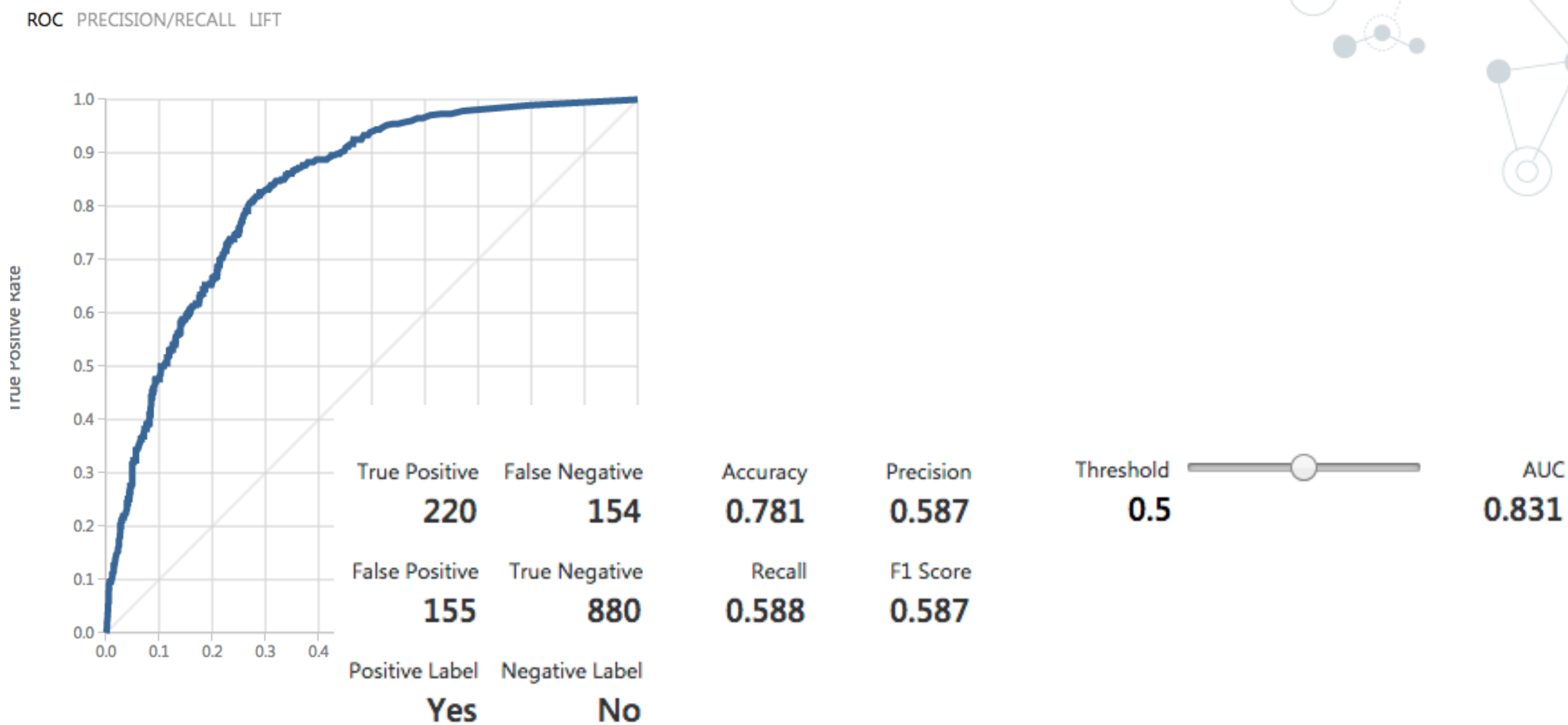
tenure

PREDICATE tenure ≤ 70.5

SPLIT GAIN 0.425254

GAIN P-VALUE 0.00116

Evaluate Model



Feature Importance

Feature	Score
TotalCharges	0.046842
tenure	0.044713
Contract	0.02626
InternetService	0.014904
MonthlyCharges	0.012775
gender	0.011356
OnlineSecurity	0.007807
StreamingTV	0.006388
MultipleLines	0.005678
DeviceProtection	0.004968
PaperlessBilling	0.004968
OnlineBackup	0.003549
StreamingMovies	0.003549
Partner	0.002129
PaymentMethod	0.002129
customerID	0
Dependents	0
SeniorCitizen	-0.00071
PhoneService	-0.004258
TechSupport	-0.007097

Overfitting?

Fold Number	Number of examples in fold	Model	Accuracy	Precision	Recall	F-Score	AUC	Average Log Loss	Training Log Loss
0	563	FastTree (Boosted Trees) Classification	0.763766	0.562914	0.559211	0.561056	0.839192	0.534286	8.393712
1	563	FastTree (Boosted Trees) Classification	0.813499	0.6	0.576923	0.588235	0.845248	0.50832	5.930884
2	563	FastTree (Boosted Trees) Classification	0.795737	0.570423	0.6	0.584838	0.825926	0.566742	-2.887989
3	564	FastTree (Boosted Trees) Classification	0.769504	0.597315	0.559748	0.577922	0.824862	0.61291	-3.05289
4	563	FastTree (Boosted Trees) Classification	0.786856	0.57764	0.641379	0.607843	0.82267	0.604139	-5.900671
5	563	FastTree (Boosted Trees) Classification	0.797513	0.693878	0.596491	0.641509	0.821645	0.632643	-3.037724
6	564	FastTree (Boosted Trees) Classification	0.801418	0.641892	0.616883	0.629139	0.860722	0.51324	12.456428
7	564	FastTree (Boosted Trees) Classification	0.776596	0.518519	0.534351	0.526316	0.806322	0.596914	-10.130905
8	563	FastTree (Boosted Trees) Classification	0.783304	0.598639	0.582781	0.590604	0.813661	0.67963	-16.882113
9	564	FastTree (Boosted Trees) Classification	0.771277	0.648438	0.497006	0.562712	0.837252	0.576799	5.057164
Mean	5634	FastTree (Boosted Trees) Classification	0.785947	0.600966	0.576477	0.587017	0.82975	0.582562	-1.00541
Standard Deviation	5634	FastTree (Boosted Trees) Classification	0.01599	0.049781	0.041509	0.033712	0.01601	0.054216	8.979025

Overfitting?

Fold Number	Number of examples in fold	Model	Accuracy	Precision	Recall	F-Score	AUC	Average Log Loss	Training Log Loss
0	563	FastTree (Boosted Trees) Classification	0.763766	0.562914	0.559211	0.561056	0.839192	0.534286	8.393712
1	563	FastTree (Boosted Trees) Classification	0.813499	0.6	0.576923	0.588235	0.845248	0.50832	5.930884
2	563	FastTree (Boosted Trees) Classification	0.795737	0.570423	0.6	0.584838	0.825926	0.566742	-2.887989
3	564	FastTree (Boosted Trees) Classification	0.769504	0.597315	0.559748	0.577922	0.824862	0.61291	-3.05289
4	563	FastTree (Boosted Trees) Classification	0.786856	0.57764	0.641379	0.607843	0.82267	0.604139	-5.900671
5	563	FastTree (Boosted Trees) Classification	0.797513	0.693878	0.596491	0.641509	0.821645	0.632643	-3.037724
6	564	FastTree (Boosted Trees) Classification	0.801418	0.641892	0.616883	0.629139	0.860722	0.51324	12.456428
7	564	FastTree (Boosted Trees) Classification	0.776596	0.518519	0.534351	0.526316	0.806322	0.596914	-10.130905
8	563	FastTree (Boosted Trees) Classification	0.783304	0.598639	0.582781	0.590604	0.813661	0.67963	-16.882113
9	564	FastTree (Boosted Trees) Classification	0.771277	0.648438	0.497006	0.562712	0.837252	0.576799	5.057164
Mean	5634	FastTree (Boosted Trees) Classification	0.785947	0.600966	0.576477	0.587017	0.82975	0.582562	-1.00541
Standard Deviation	5634	FastTree (Boosted Trees) Classification	0.01599	0.049781	0.041509	0.033712	0.01601	0.054216	8.979025

Overfitting?

Fold Number	Number of examples in fold	Model	Accuracy	Precision	Recall	F-Score	AUC	Average Log Loss	Training Log Loss
0	563	FastTree (Boosted Trees) Classification	0.763766	0.562914	0.559211	0.561056	0.839192	0.534286	8.393712
1	563	FastTree (Boosted Trees) Classification	0.813499	0.6	0.576923	0.588235	0.845248	0.50832	5.930884
2	563	FastTree (Boosted Trees) Classification	0.795737	0.570423	0.6	0.584838	0.825926	0.566742	-2.887989
3	564	FastTree (Boosted Trees) Classification	0.769504	0.597315	0.559748	0.577922	0.824862	0.61291	-3.05289
4	563	FastTree (Boosted Trees) Classification	0.786856	0.57764	0.641379	0.607843	0.82267	0.604139	-5.900671
5	563	FastTree (Boosted Trees) Classification	0.797513	0.693878	0.596491	0.641509	0.821645	0.632643	-3.037724
6	564	FastTree (Boosted Trees) Classification	0.801418	0.641892	0.616883	0.629139	0.860722	0.51324	12.456428
7	564	FastTree (Boosted Trees) Classification	0.776596	0.518519	0.534351	0.526316	0.806322	0.596914	-10.130905
8	563	FastTree (Boosted Trees) Classification	0.783304	0.598639	0.582781	0.590604	0.813661	0.67963	-16.882113
9	564	FastTree (Boosted Trees) Classification	0.771277	0.648438	0.497006	0.562712	0.837252	0.576799	5.057164
Mean	5634	FastTree (Boosted Trees) Classification	0.785947	0.600966	0.576477	0.587017	0.82975	0.582562	-1.00541
Standard Deviation	5634	FastTree (Boosted Trees) Classification	0.01599	0.049781	0.041509	0.033712	0.01601	0.054216	8.979025

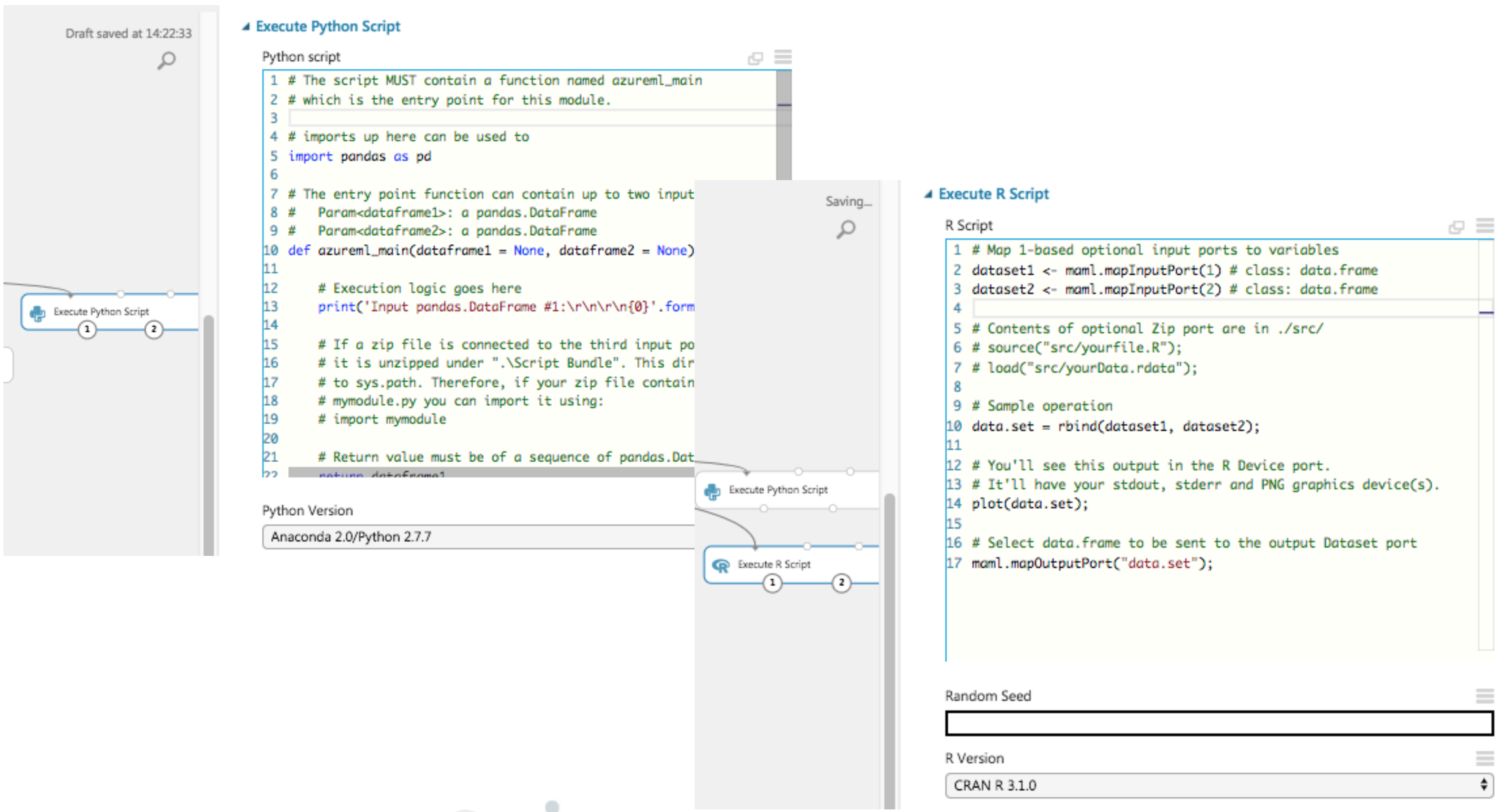
Min: $-3s+x = 0.73$

Overfitting?

Fold Number	Number of examples in fold	Model	Accuracy	Precision	Recall	F-Score	AUC	Average Log Loss	Training Log Loss
0	563	FastTree (Boosted Trees) Classification	0.763766	0.562914	0.559211	0.561056	0.839192	0.534286	8.393712
1	563	FastTree (Boosted Trees) Classification	0.813499	0.6	0.576923	0.588235	0.845248	0.50832	5.930884
2	563	FastTree (Boosted Trees) Classification	0.795737	0.570423	0.6	0.584838	0.825926	0.566742	-2.887989
3	564	FastTree (Boosted Trees) Classification	0.769504	0.597315	0.559748	0.577922	0.824862	0.61291	-3.05289
4	563	FastTree (Boosted Trees) Classification	0.786856	0.57764	0.641379	0.607843	0.82267	0.604139	-5.900671
5	563	FastTree (Boosted Trees) Classification	0.797513	0.693878	0.596491	0.641509	0.821645	0.632643	-3.037724
6	564	FastTree (Boosted Trees) Classification	0.801418	0.641892	0.616883	0.629139	0.860722	0.51324	12.456428
7	564	FastTree (Boosted Trees) Classification	0.776596	0.518519	0.534351	0.526316	0.806322	0.596914	-10.130905
8	563	FastTree (Boosted Trees) Classification	0.783304	0.598639	0.582781	0.590604	0.813661	0.67963	-16.882113
9	564	FastTree (Boosted Trees) Classification	0.771277	0.648438	0.497006	0.562712	0.837252	0.576799	5.057164
Mean	5634	FastTree (Boosted Trees) Classification	0.785947	0.600966	0.576477	0.587017	0.82975	0.582562	-1.00541
Standard Deviation	5634	FastTree (Boosted Trees) Classification	0.01599	0.049781	0.041509	0.033712	0.01601	0.054216	8.979025

Min: -3s+x = 0.73
Max: 3s+x = 0.83

I like to get my hands dirty…



Play time!

Thanks!

Any questions?