Open for Innovation®
KNIME

Guided Analytics for Machine Learning Automation

Christian Dietz, Simon Schmid

KNIME

© 2018 KNIME AG. All rights reserved.

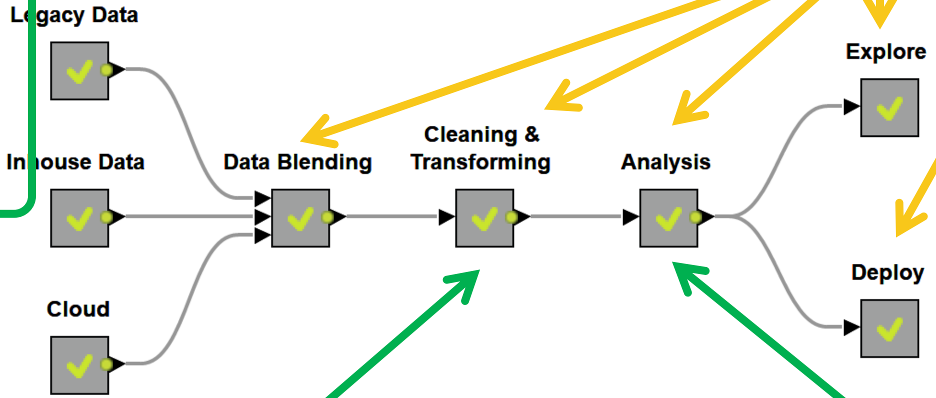
Automating Everything?

Human Input Needed!

- Data Selection – Is this relevant?!
- Analysis Goal – What is interesting?
- Exploration – This looks weird?...

Automating Data Integration:

- Parsing
- Record Matching
- ...



Automating Data Proc:

- Feature Selection
- Feature Construction
- Data Cleaning

Automating Analytics:

- Parameter Optimization
- Model Selection
- Ensemble Construction

© 2018 KNIME AG. All rights reserved.

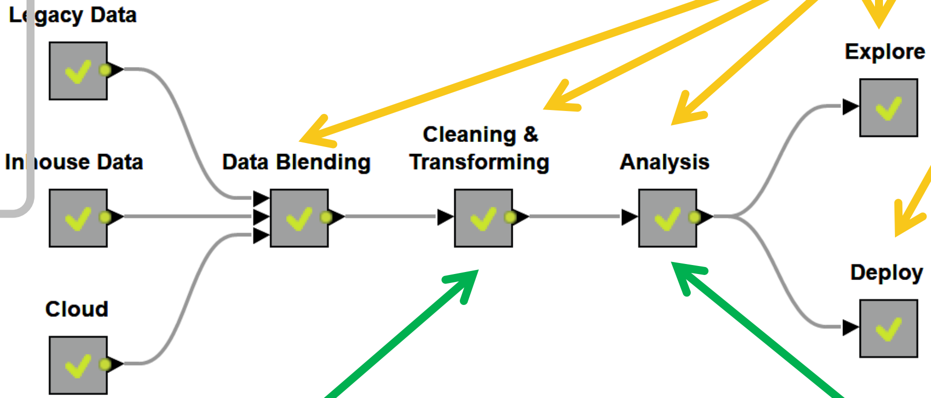
Automating Everything?

Human Input Needed!

- Data Selection – Is this relevant?!
- Analysis Goal – What is interesting?
- Exploration – This looks weird?...

Automating Data Integration:

- Parsing
- Record Matching
- …



Automating Data Proc:

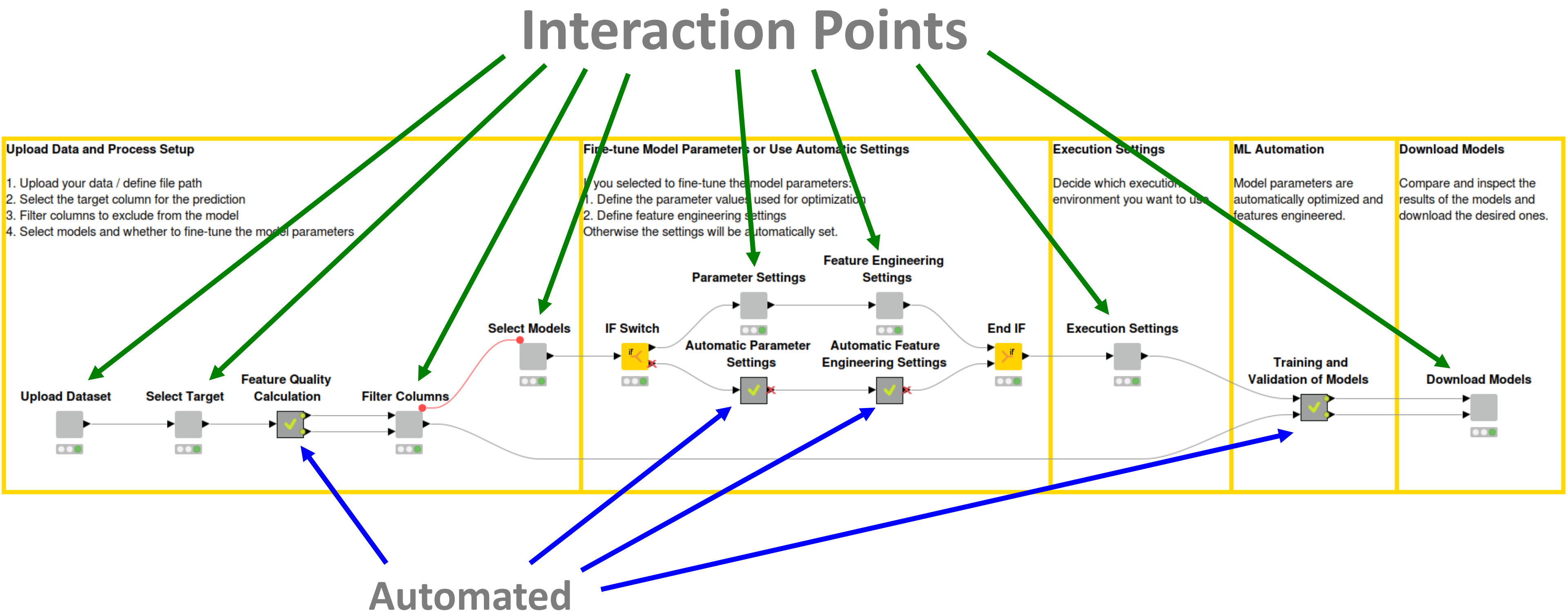
- Feature Selection
- Feature Construction
- Data Cleaning

Automating Analytics:

- Parameter Optimization
- Model Selection
- Ensemble Construction

© 2018 KNIME AG. All rights reserved.

Building a Guided Automation Workflow



Upload Data and Process Setup

1. Upload your data / define file path
2. Select the target column for the prediction
3. Filter columns to exclude from the model
4. Select models and whether to fine-tune the model parameters

Fine-tune Model Parameters or Use Automatic Settings

If you selected to fine-tune the model parameters:
1. Define the parameter values used for optimization
2. Define feature engineering settings
Otherwise the settings will be automatically set.

Execution Settings

Decide which execution environment you want to use.

ML Automation

Model parameters are automatically optimized and features engineered.

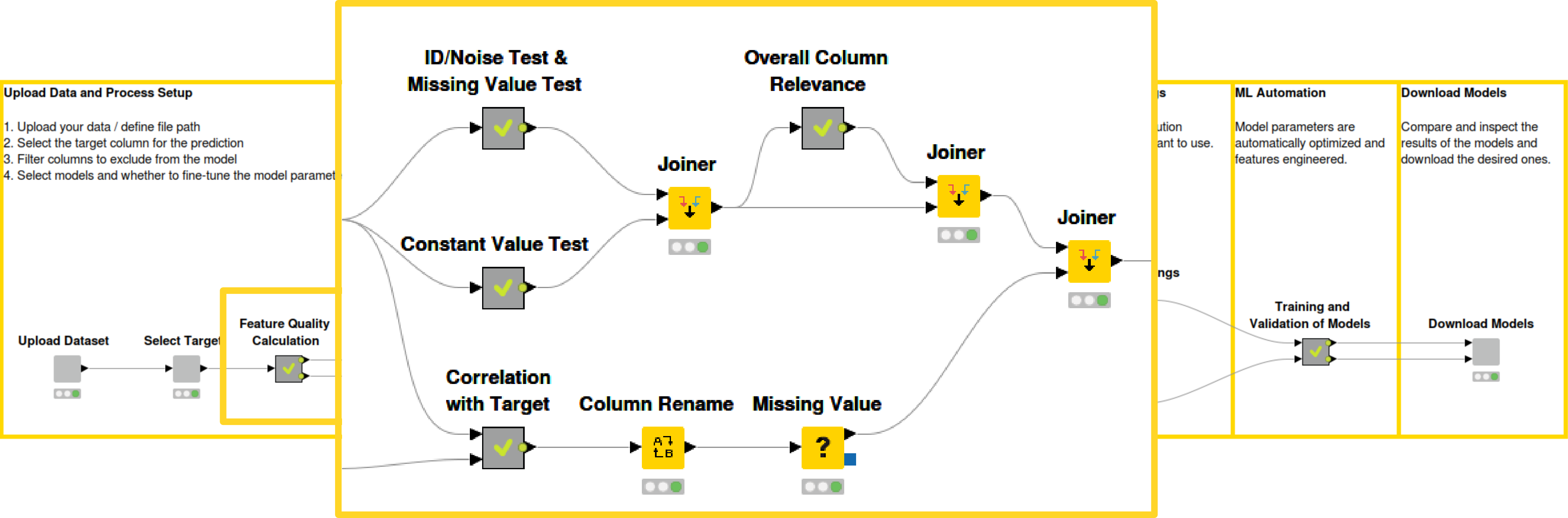
Download Models

Compare and inspect the results of the models and download the desired ones.

Upload Dataset
Select Target
Feature Quality Calculation
Filter Columns
Select Models
IF Switch
Parameter Settings
Feature Engineering Settings
Automatic Parameter Settings
Automatic Feature Engineering Settings
End IF
Execution Settings
Training and Validation of Models
Download Models
Automated

© 2018 KNIME AG. All rights reserved.

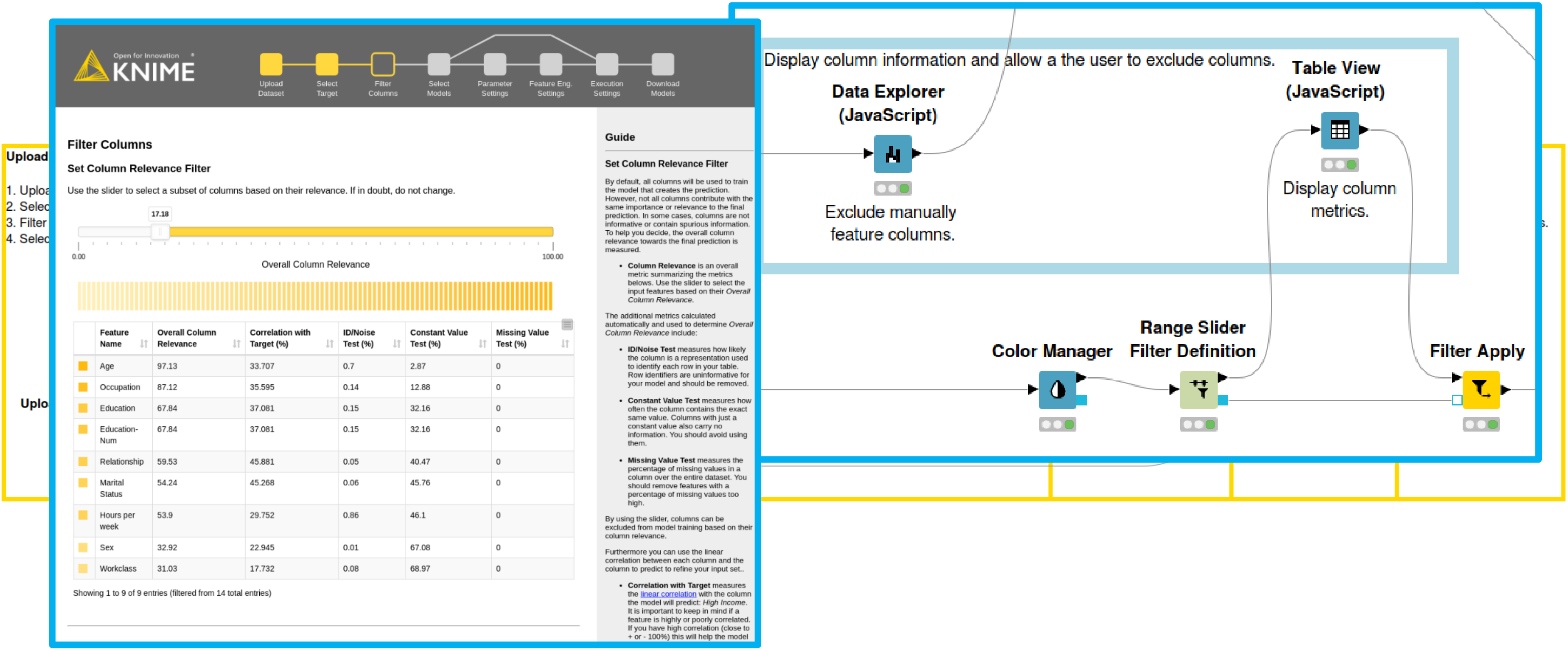
Guided Automation: Automation + Interaction



Upload Data and Process Setup
1. Upload your data / define file path
2. Select the target column for the prediction
3. Filter columns to exclude from the model
4. Select models and whether to fine-tune the model paramet
Upload Dataset
Select Target
Feature Quality Calculation
ID/Noise Test & Missing Value Test
Overall Column Relevance
Joiner
Joiner
Joiner
Constant Value Test
Correlation with Target
Column Rename
Missing Value
ıs
ıtion
ant to use.
ngs
ML Automation
Model parameters are automatically optimized and features engineered.
Training and Validation of Models
Download Models
Compare and inspect the results of the models and download the desired ones.
Download Models

© 2018 KNIME AG. All rights reserved.

Guided Automation: Automation + Interaction



Filter Columns

Set Column Relevance Filter

Use the slider to select a subset of columns based on their relevance. If in doubt, do not change.

	Feature Name	Overall Column Relevance	Correlation with Target (%)	ID/Noise Test (%)	Constant Value Test (%)	Missing Value Test (%)
🟨	Age	97.13	33.707	0.7	2.87	0
🟨	Occupation	87.12	35.595	0.14	12.88	0
🟨	Education	67.84	37.081	0.15	32.16	0
🟨	Education-Num	67.84	37.081	0.15	32.16	0
🟨	Relationship	59.53	45.881	0.05	40.47	0
🟨	Marital Status	54.24	45.268	0.06	45.76	0
🟨	Hours per week	53.9	29.752	0.86	46.1	0
🟨	Sex	32.92	22.945	0.01	67.08	0
🟨	Workclass	31.03	17.732	0.08	68.97	0

Showing 1 to 9 of 9 entries (filtered from 14 total entries)

Guide

Set Column Relevance Filter

By default, all columns will be used to train the model that creates the prediction. However, not all columns contribute with the same importance or relevance to the final prediction. In some cases, columns are not informative or contain spurious information. To help you decide, the overall column relevance towards the final prediction is measured.

- **Column Relevance** is an overall metric summarizing the metrics belows. Use the slider to select the input features based on their *Overall Column Relevance*.

The additional metrics calculated automatically and used to determine *Overall Column Relevance* include:

- **ID/Noise Test** measures how likely the column is a representation used to identify each row in your table. Row identifiers are uninformative for your model and should be removed.
- **Constant Value Test** measures how often the column contains the exact same value. Columns with just a constant value also carry no information. You should avoid using them.
- **Missing Value Test** measures the percentage of missing values in a column over the entire dataset. You should remove features with a percentage of missing values too high.

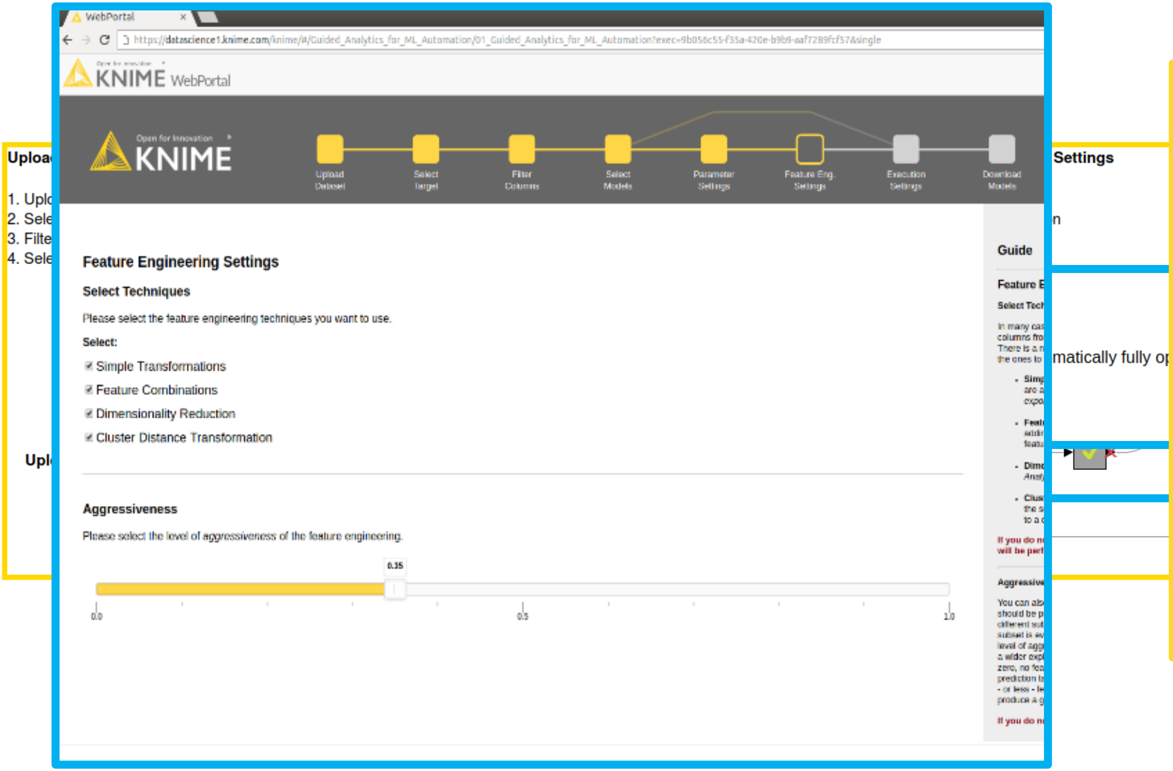
By using the slider, columns can be excluded from model training based on their column relevance.

Furthermore you can use the linear correlation between each column and the column to predict to refine your input set..

- **Correlation with Target** measures the linear correlation with the column the model will predict: *High Income*. It is important to keep in mind if a feature is highly or poorly correlated. If you have high correlation (close to + or - 100%) this will help the model

© 2018 KNIME AG. All rights reserved.

Guided Automation: Automation + Interaction



Feature Engineering Settings

Select Techniques

Please select the feature engineering techniques you want to use.

Select:

☑ Simple Transformations

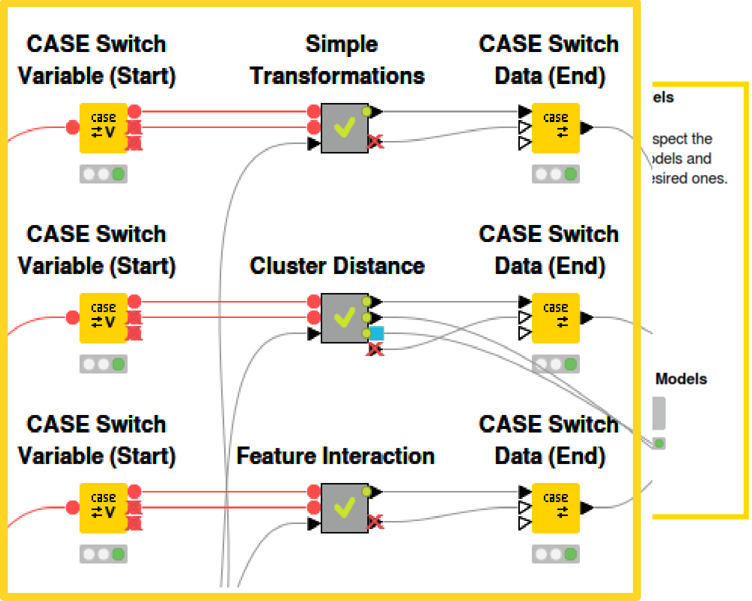
☑ Feature Combinations

☑ Dimensionality Reduction

☑ Cluster Distance Transformation

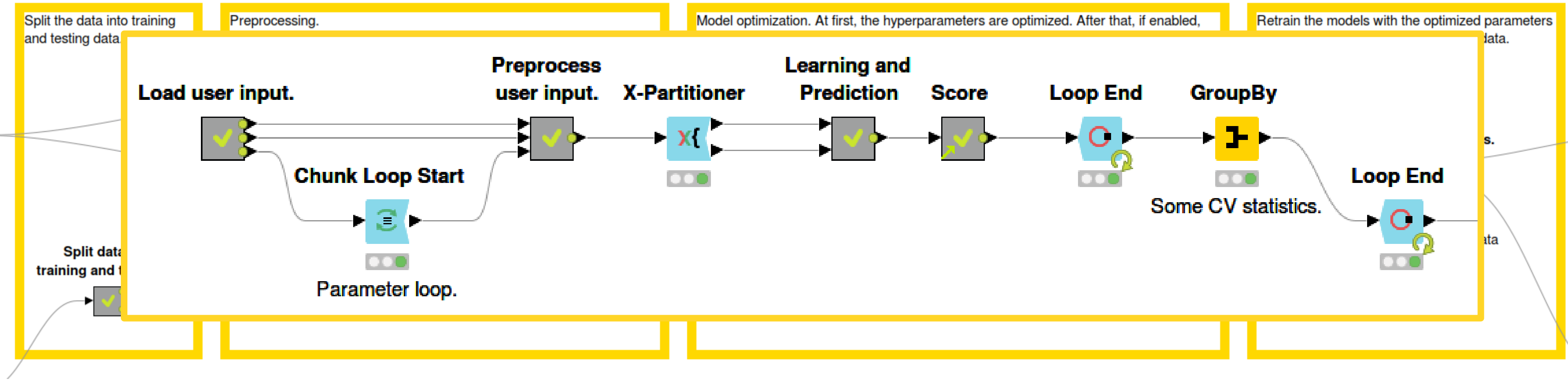
Aggressiveness

Please select the level of aggressiveness of the feature engineering.



© 2018 KNIME AG. All rights reserved.

Guided Automation: Automation + Interaction



Split the data into training and testing data.

Preprocessing.

Model optimization. At first, the hyperparameters are optimized. After that, if enabled,

Retrain the models with the optimized parameters

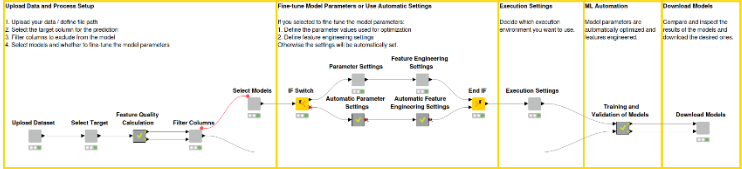
Some CV statistics.

Parameter loop.

© 2018 KNIME AG. All rights reserved.

Guided Automation: Automation + Interaction

Workflow ✓



Upload Data and Process Setup

1. Upload your data / define file path
2. Select the target column for the prediction
3. Filter columns to exclude from the model
4. Select models and whether to fine-tune the model parameters

Fine-tune Model Parameters or Use Automatic Settings

If you selected to fine-tune the model parameters:
1. Define the parameter values used for optimization
2. Define feature engineering settings
Otherwise the settings will be automatically set.

Execution Settings

Decide which execution environment you want to use.

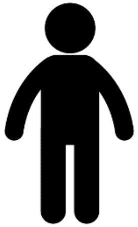
ML Automation

Model parameters are automatically optimized and features engineered.

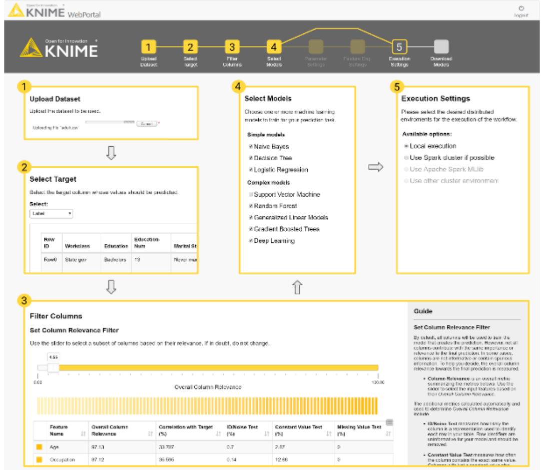
Download Models

Compare and inspect the results of the models and download the desired ones.

Upload Dataset
Select Target
Feature Quality Calculation
Filter Columns
Select Models
IF Switch
Parameter Settings
Automatic Parameter Settings
Feature Engineering Settings
Automatic Feature Engineering Settings
End IF
Execution Settings
Training and Validation of Models
Download Models



WebPortal



1 **Upload Dataset**

Upload the dataset to be used.

Uploading File "adult.csv"

2 **Select Target**

Select the target column whose values should be predicted.

Select: Label

Row ID	Workclass	Education	Education-Num	Marital St
Row0	State gov	Bachelors	13	Never mar

3 **Filter Columns**

Set Column Relevance Filter

Use the slider to select a subset of columns based on their relevance. If in doubt, do not change.

	Feature Name	Overall Column Relevance	Correlation with Target (%)	ID/Noise Test (%)	Constant Value Test (%)	Missing Value Test (%)
🟨	Age	87.13	33.707	0.7	2.87	0
🟨	Occupation	87.12	35.606	0.14	12.68	0

Guide

Set Column Relevance Filter

By default, all columns will be used to train the model that creates the prediction. However, not all columns contribute with the same importance or relevance to the final prediction. In some cases, columns are not informative or contain spurious information. To help you decide, the overall column relevance towards the final prediction is measured.

- **Column Relevance** is an overall metric summarizing the metrics below. Use the slider to select the input features based on their Overall Column Relevance.

The additional metrics calculated automatically and used to determine Overall Column Relevance include:

- **ID/Noise Test** measures how likely the column is a representation used to identify each row in your table. Row identifiers are uninformative for your model and should be removed.
- **Constant Value Test** measures how often the column contains the exact same value.

4 **Select Models**

Choose one or more machine learning models to train for your prediction task.

Simple models

☑ Naive Bayes
☑ Decision Tree
☑ Logistic Regression

Complex models

□ Support Vector Machine
☑ Random Forest
☑ Generalized Linear Models
☑ Gradient Boosted Trees
☑ Deep Learning

5 **Execution Settings**

Please select the desired distributed enviroments for the execution of the workflow.

Available options:

● Local execution
○ Use Spark cluster if possible
○ Use Apache Spark MLlib
○ Use other cluster environment

KNIME Server

© 2018 KNIME AG. All rights reserved.

Guided Automation on KNIME Server

Live Demo

© 2018 KNIME AG. All rights reserved.

Scoring Workflow

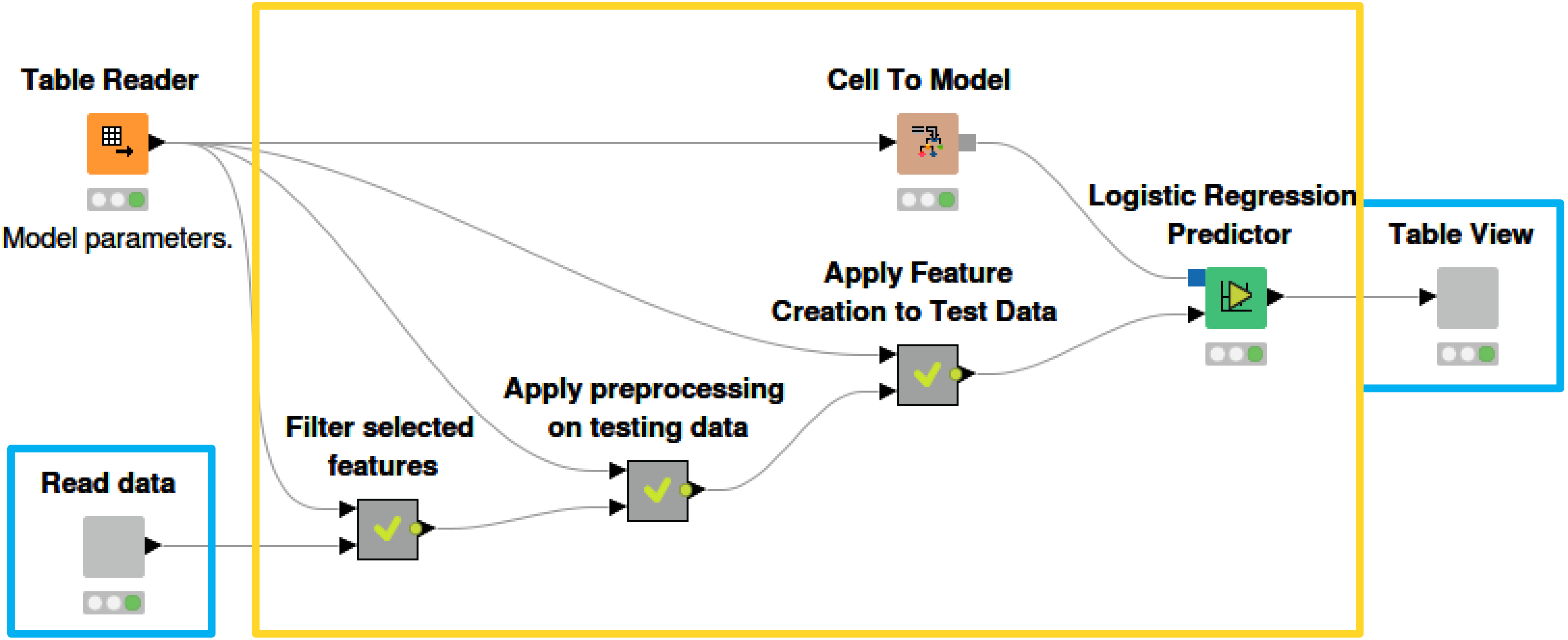


Table Reader
Model parameters.
Cell To Model
Logistic Regression Predictor
Table View
Apply Feature Creation to Test Data
Apply preprocessing on testing data
Filter selected features
Read data

© 2018 KNIME AG. All rights reserved.

Scoring Workflow

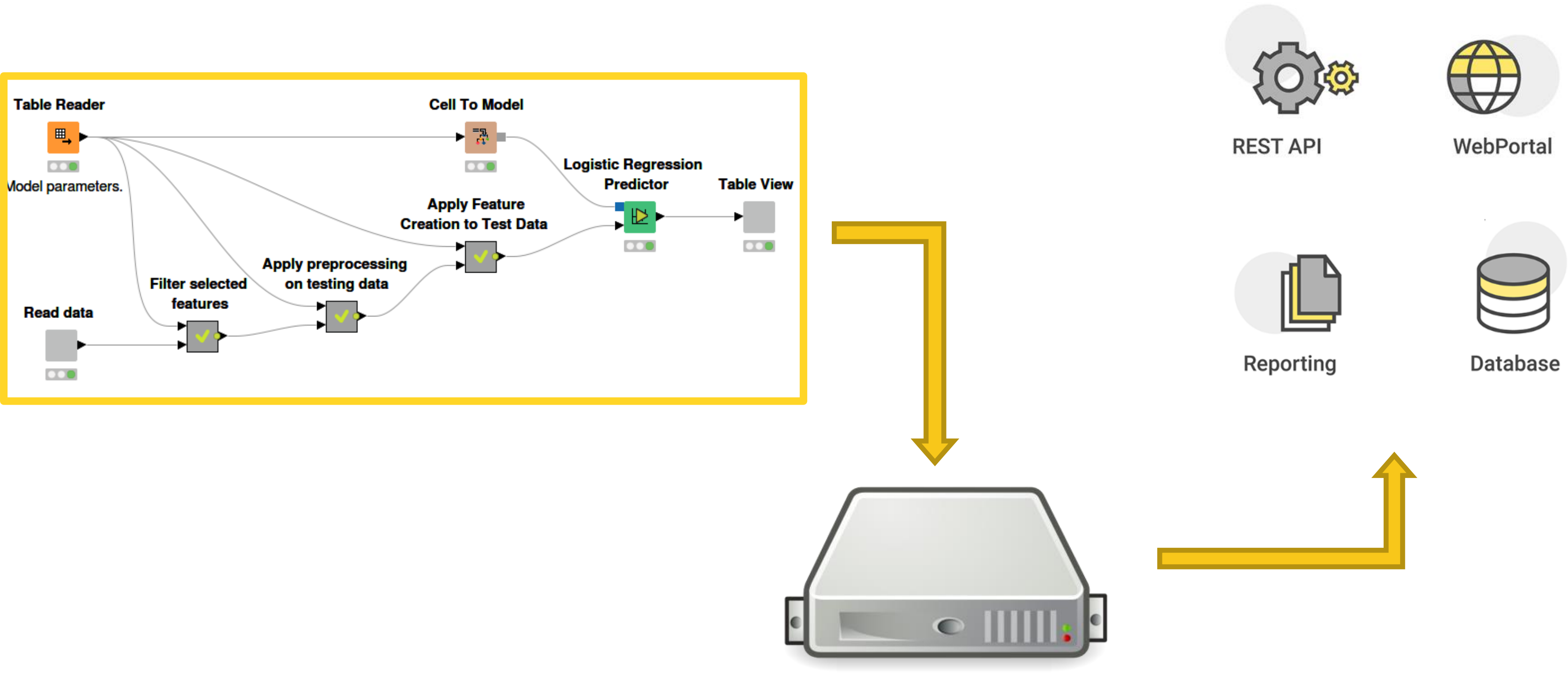


Table Reader
Model parameters.
Cell To Model
Logistic Regression Predictor
Table View
Apply Feature Creation to Test Data
Apply preprocessing on testing data
Filter selected features
Read data

REST API

WebPortal

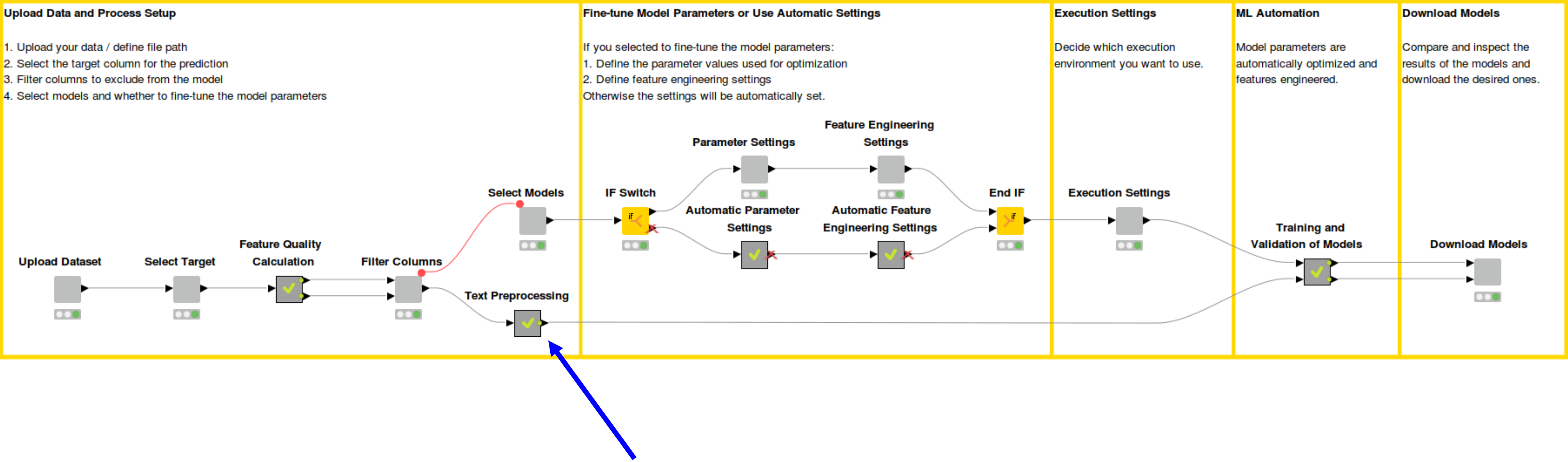
Reporting

Database

KNIME Server

© 2018 KNIME AG. All rights reserved.

Customize the Blueprint for Text Processing



Upload Data and Process Setup

1. Upload your data / define file path
2. Select the target column for the prediction
3. Filter columns to exclude from the model
4. Select models and whether to fine-tune the model parameters

Fine-tune Model Parameters or Use Automatic Settings

If you selected to fine-tune the model parameters:
1. Define the parameter values used for optimization
2. Define feature engineering settings
Otherwise the settings will be automatically set.

Execution Settings

Decide which execution environment you want to use.

ML Automation

Model parameters are automatically optimized and features engineered.

Download Models

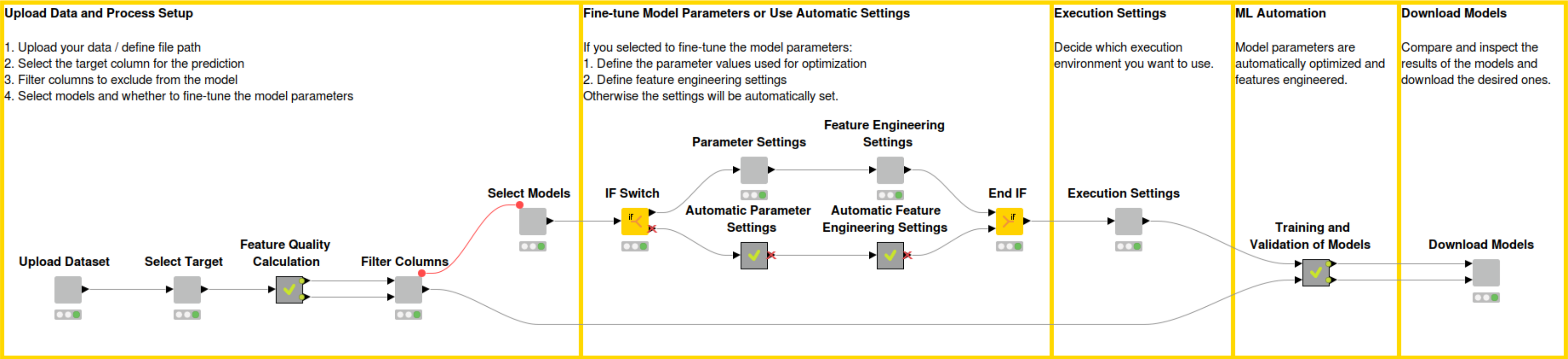
Compare and inspect the results of the models and download the desired ones.

Upload Dataset
Select Target
Feature Quality Calculation
Filter Columns
Select Models
Text Preprocessing
IF Switch
Parameter Settings
Feature Engineering Settings
Automatic Parameter Settings
Automatic Feature Engineering Settings
End IF
Execution Settings
Training and Validation of Models
Download Models

Added by in-house expert.

© 2018 KNIME AG. All rights reserved.

Thank You!



Upload Data and Process Setup

1. Upload your data / define file path
2. Select the target column for the prediction
3. Filter columns to exclude from the model
4. Select models and whether to fine-tune the model parameters

Fine-tune Model Parameters or Use Automatic Settings

If you selected to fine-tune the model parameters:
1. Define the parameter values used for optimization
2. Define feature engineering settings
Otherwise the settings will be automatically set.

Execution Settings

Decide which execution environment you want to use.

ML Automation

Model parameters are automatically optimized and features engineered.

Download Models

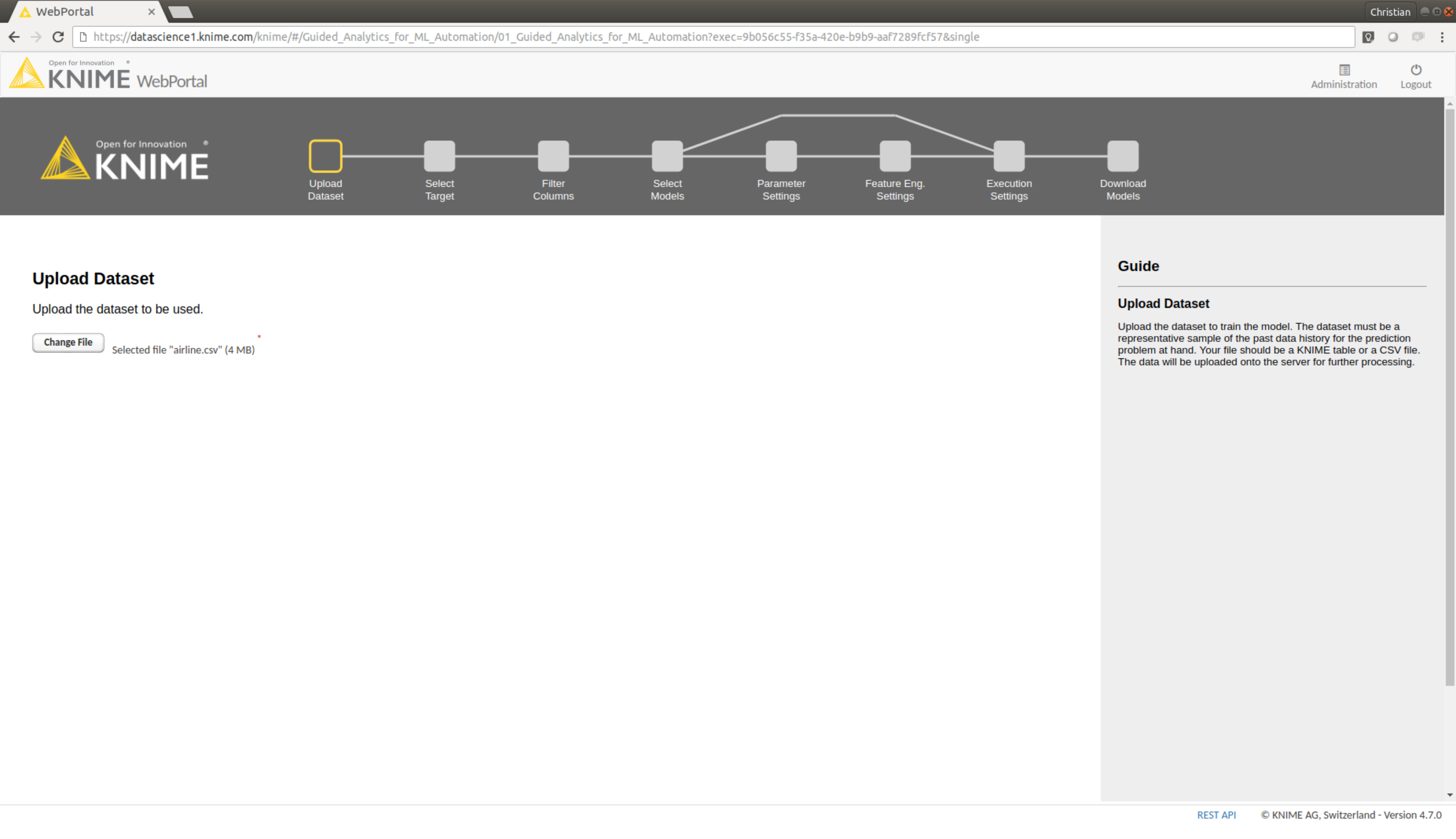
Compare and inspect the results of the models and download the desired ones.

Upload Dataset
Select Target
Feature Quality Calculation
Filter Columns
Select Models
IF Switch
Parameter Settings
Feature Engineering Settings
Automatic Parameter Settings
Automatic Feature Engineering Settings
End IF
Execution Settings
Training and Validation of Models
Download Models

Download workflow from **knime.com** and get started!

© 2018 KNIME AG. All rights reserved.

The KNIME® trademark and logo and OPEN FOR INNOVATION® trademark are used by KNIME AG under license from KNIME GmbH, and are registered in the United States. KNIME® is also registered in Germany.


WebPortal
Christian
https://datascience1.knime.com/knime/#/Guided_Analytics_for_ML_Automation/01_Guided_Analytics_for_ML_Automation?exec=9b056c55-f35a-420e-b9b9-aaf7289fcf57&single
Open for Innovation KNIME WebPortal
Administration
Logout
Open for Innovation KNIME
Upload Dataset
Select Target
Filter Columns
Select Models
Parameter Settings
Feature Eng. Settings
Execution Settings
Download Models

Upload Dataset

Upload the dataset to be used.

Change File Selected file "airline.csv" (4 MB)

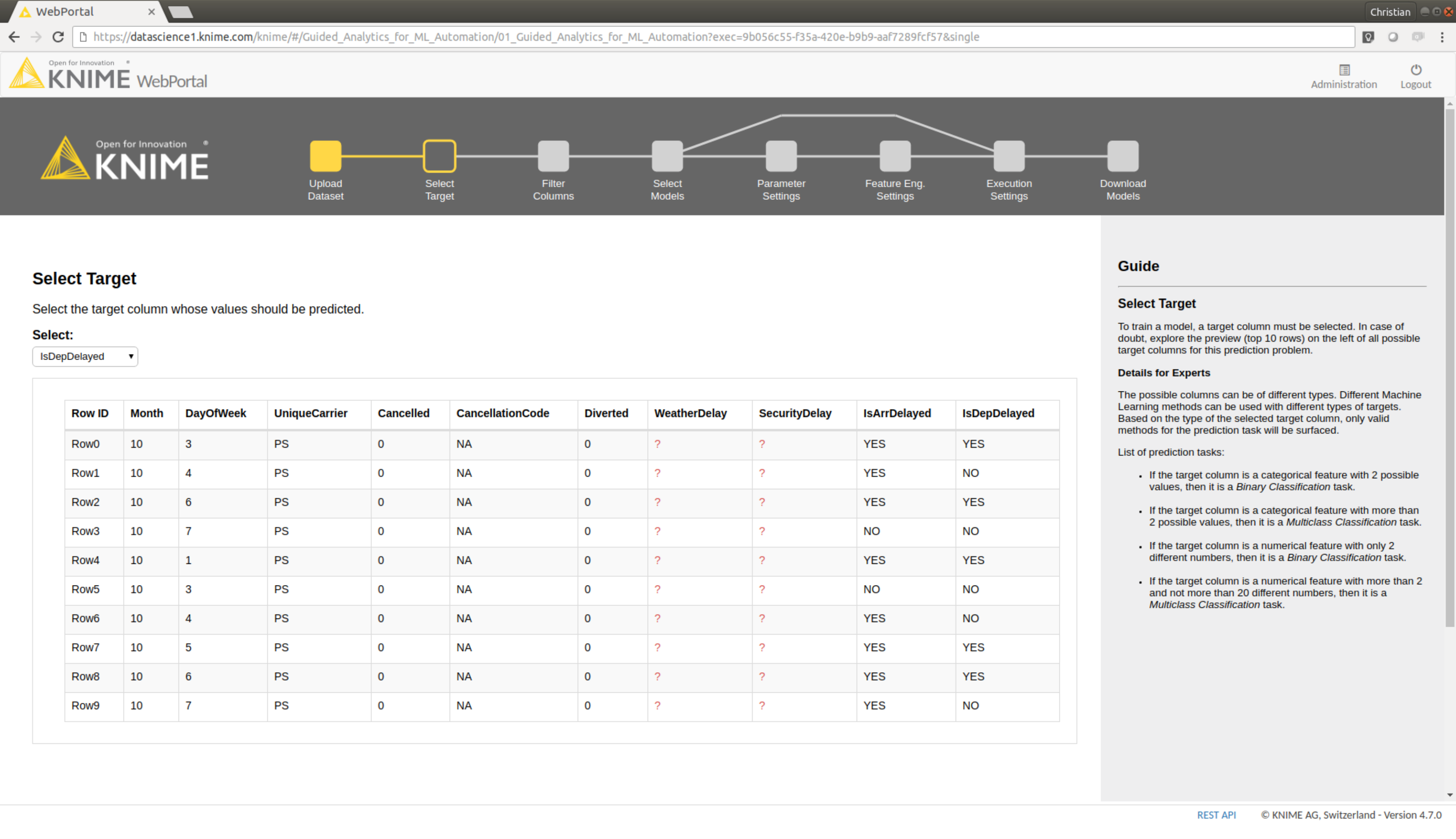
Guide

Upload Dataset

Upload the dataset to train the model. The dataset must be a representative sample of the past data history for the prediction problem at hand. Your file should be a KNIME table or a CSV file. The data will be uploaded onto the server for further processing.

REST API © KNIME AG, Switzerland - Version 4.7.0

© 2018 KNIME AG. All rights reserved.



Select Target

Select the target column whose values should be predicted.

Select: IsDepDelayed ▾

Row ID	Month	DayOfWeek	UniqueCarrier	Cancelled	CancellationCode	Diverted	WeatherDelay	SecurityDelay	IsArrDelayed	IsDepDelayed
Row0	10	3	PS	0	NA	0	?	?	YES	YES
Row1	10	4	PS	0	NA	0	?	?	YES	NO
Row2	10	6	PS	0	NA	0	?	?	YES	YES
Row3	10	7	PS	0	NA	0	?	?	NO	NO
Row4	10	1	PS	0	NA	0	?	?	YES	YES
Row5	10	3	PS	0	NA	0	?	?	NO	NO
Row6	10	4	PS	0	NA	0	?	?	YES	NO
Row7	10	5	PS	0	NA	0	?	?	YES	YES
Row8	10	6	PS	0	NA	0	?	?	YES	YES
Row9	10	7	PS	0	NA	0	?	?	YES	NO

Guide

Select Target

To train a model, a target column must be selected. In case of doubt, explore the preview (top 10 rows) on the left of all possible target columns for this prediction problem.

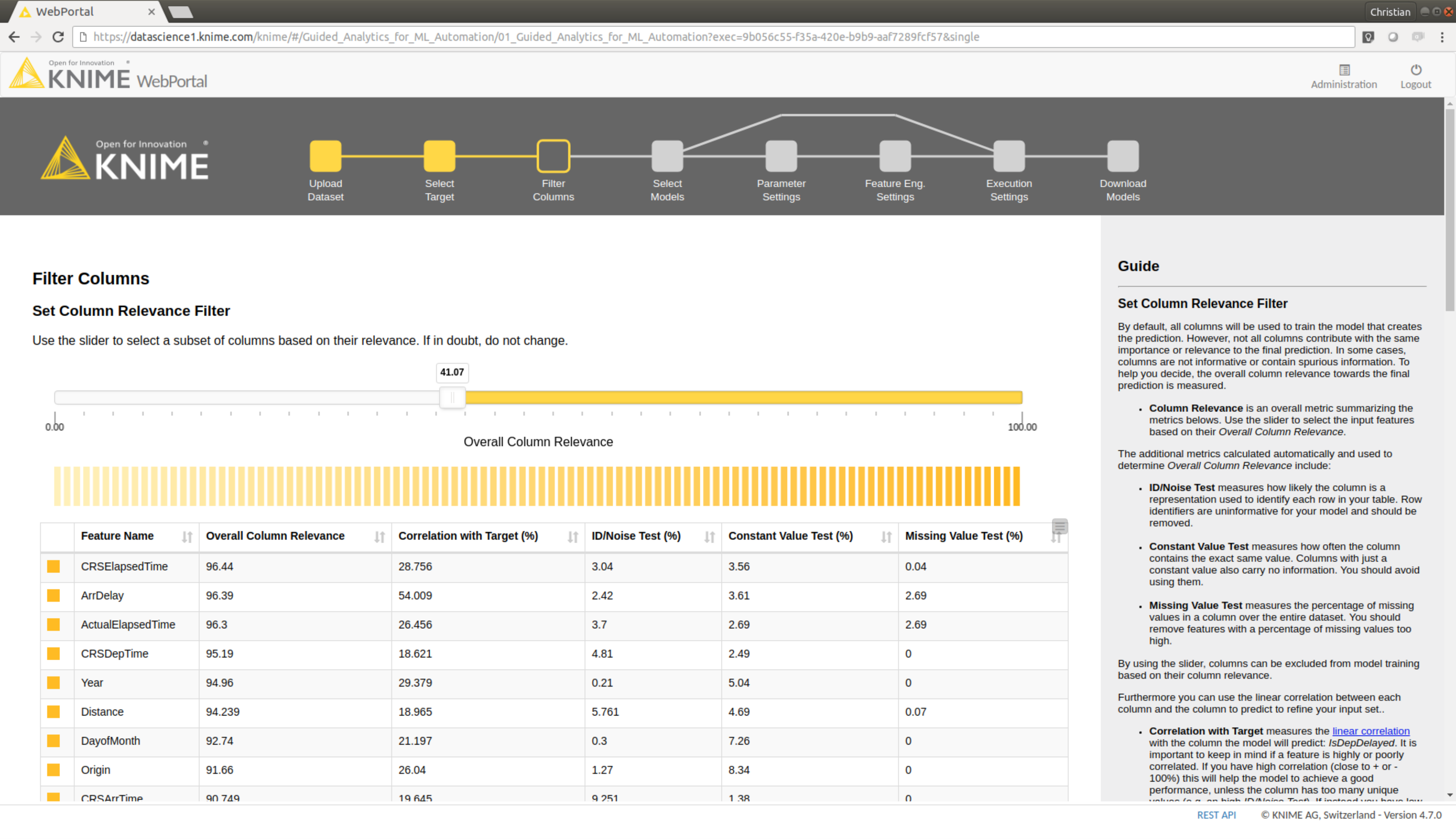
Details for Experts

The possible columns can be of different types. Different Machine Learning methods can be used with different types of targets. Based on the type of the selected target column, only valid methods for the prediction task will be surfaced.

List of prediction tasks:

- If the target column is a categorical feature with 2 possible values, then it is a *Binary Classification* task.
- If the target column is a categorical feature with more than 2 possible values, then it is a *Multiclass Classification* task.
- If the target column is a numerical feature with only 2 different numbers, then it is a *Binary Classification* task.
- If the target column is a numerical feature with more than 2 and not more than 20 different numbers, then it is a *Multiclass Classification* task.

© 2018 KNIME AG. All rights reserved.


WebPortal
Christian
https://datascience1.knime.com/knime/#/Guided_Analytics_for_ML_Automation/01_Guided_Analytics_for_ML_Automation?exec=9b056c55-f35a-420e-b9b9-aaf7289fcf57&single
Open for Innovation KNIME WebPortal
Administration
Logout
Open for Innovation KNIME
Upload Dataset
Select Target
Filter Columns
Select Models
Parameter Settings
Feature Eng. Settings
Execution Settings
Download Models
41.07
0.00
100.00
Overall Column Relevance
REST API
© KNIME AG, Switzerland - Version 4.7.0

Filter Columns

Set Column Relevance Filter

Use the slider to select a subset of columns based on their relevance. If in doubt, do not change.

	Feature Name	Overall Column Relevance	Correlation with Target (%)	ID/Noise Test (%)	Constant Value Test (%)	Missing Value Test (%)
🟧	CRSElapsedTime	96.44	28.756	3.04	3.56	0.04
🟧	ArrDelay	96.39	54.009	2.42	3.61	2.69
🟧	ActualElapsedTime	96.3	26.456	3.7	2.69	2.69
🟧	CRSDepTime	95.19	18.621	4.81	2.49	0
🟧	Year	94.96	29.379	0.21	5.04	0
🟧	Distance	94.239	18.965	5.761	4.69	0.07
🟧	DayofMonth	92.74	21.197	0.3	7.26	0
🟧	Origin	91.66	26.04	1.27	8.34	0
🟧	CRSArrTime	90.749	19.645	9.251	1.38	0

Guide

Set Column Relevance Filter

By default, all columns will be used to train the model that creates the prediction. However, not all columns contribute with the same importance or relevance to the final prediction. In some cases, columns are not informative or contain spurious information. To help you decide, the overall column relevance towards the final prediction is measured.

- **Column Relevance** is an overall metric summarizing the metrics belows. Use the slider to select the input features based on their *Overall Column Relevance*.

The additional metrics calculated automatically and used to determine *Overall Column Relevance* include:

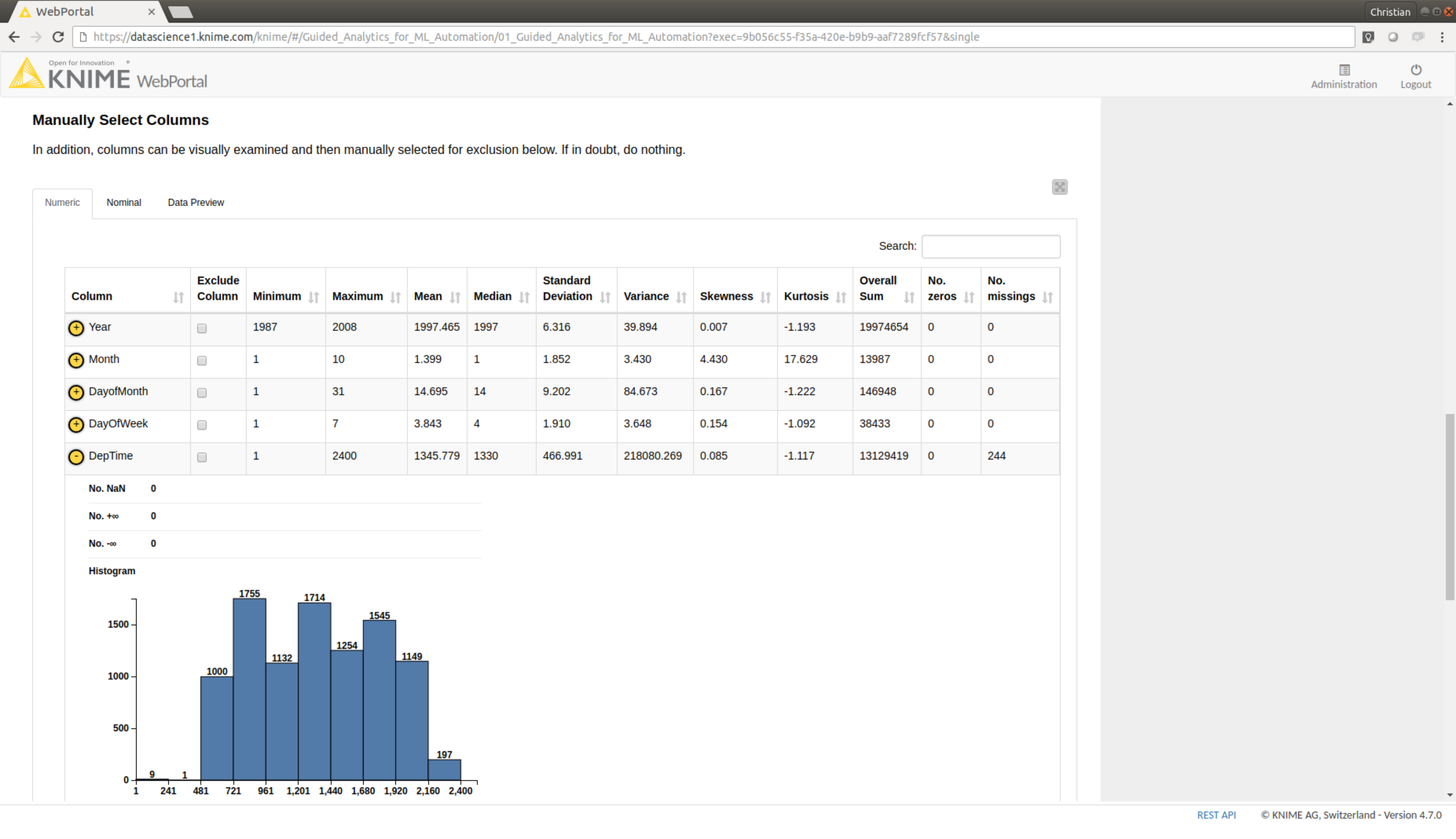
- **ID/Noise Test** measures how likely the column is a representation used to identify each row in your table. Row identifiers are uninformative for your model and should be removed.
- **Constant Value Test** measures how often the column contains the exact same value. Columns with just a constant value also carry no information. You should avoid using them.
- **Missing Value Test** measures the percentage of missing values in a column over the entire dataset. You should remove features with a percentage of missing values too high.

By using the slider, columns can be excluded from model training based on their column relevance.

Furthermore you can use the linear correlation between each column and the column to predict to refine your input set..

- **Correlation with Target** measures the linear correlation with the column the model will predict: *IsDepDelayed*. It is important to keep in mind if a feature is highly or poorly correlated. If you have high correlation (close to + or - 100%) this will help the model to achieve a good performance, unless the column has too many unique

© 2018 KNIME AG. All rights reserved.



WebPortal
https://datascience1.knime.com/knime/#/Guided_Analytics_for_ML_Automation/01_Guided_Analytics_for_ML_Automation?exec=9b056c55-f35a-420e-b9b9-aaf7289fcf57&single
Christian
Open for Innovation
KNIME WebPortal
Administration
Logout

Manually Select Columns

In addition, columns can be visually examined and then manually selected for exclusion below. If in doubt, do nothing.

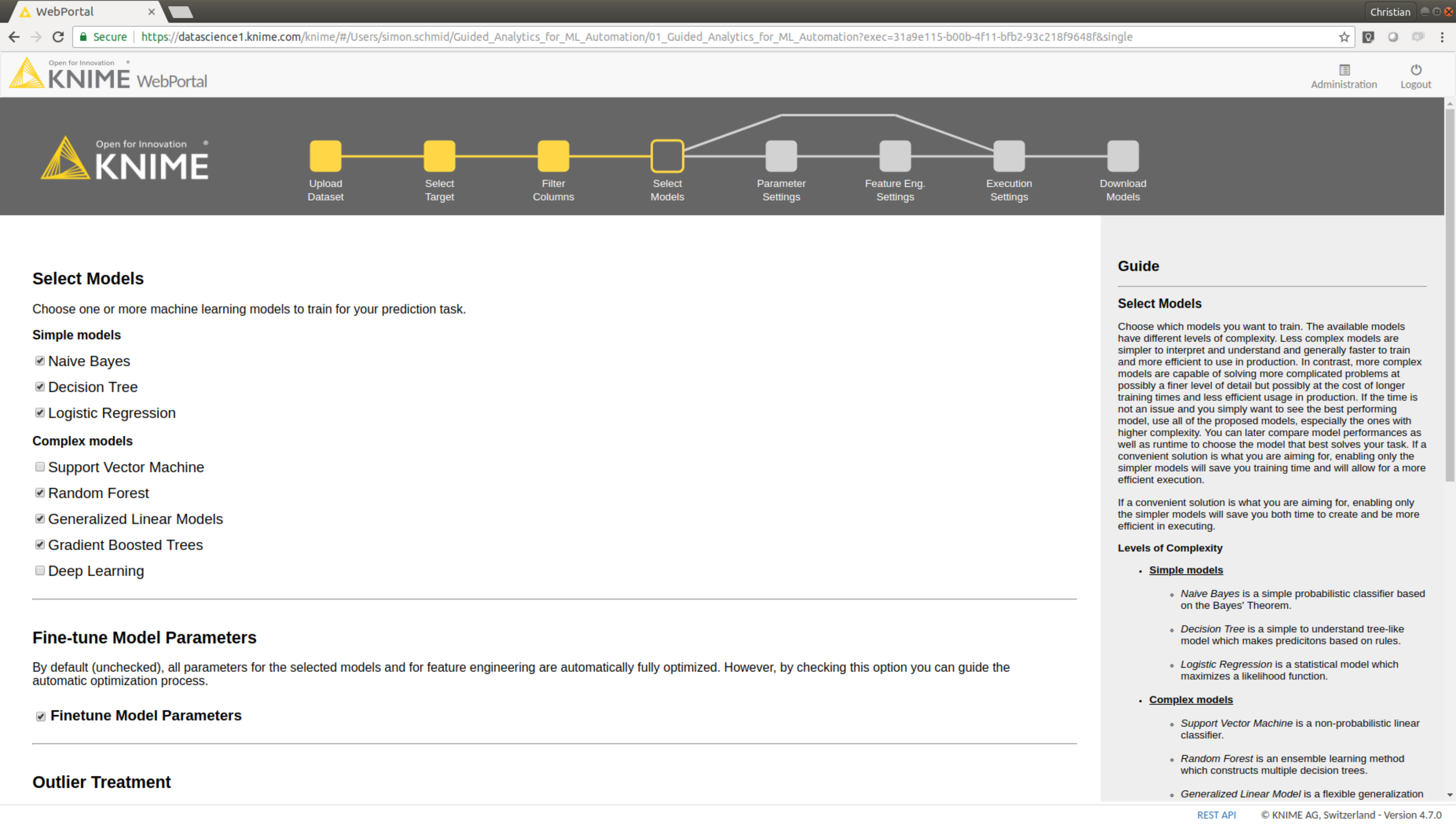
Numeric
Nominal
Data Preview
Search:

Column	Exclude Column	Minimum	Maximum	Mean	Median	Standard Deviation	Variance	Skewness	Kurtosis	Overall Sum	No. zeros	No. missings
Year	□	1987	2008	1997.465	1997	6.316	39.894	0.007	-1.193	19974654	0	0
Month	□	1	10	1.399	1	1.852	3.430	4.430	17.629	13987	0	0
DayofMonth	□	1	31	14.695	14	9.202	84.673	0.167	-1.222	146948	0	0
DayOfWeek	□	1	7	3.843	4	1.910	3.648	0.154	-1.092	38433	0	0
DepTime	□	1	2400	1345.779	1330	466.991	218080.269	0.085	-1.117	13129419	0	244

No. NaN	0
No. +∞	0
No. -∞	0

Histogram

© 2018 KNIME AG. All rights reserved.


WebPortal
Christian
Secure
https://datascience1.knime.com/knime/#/Users/simon.schmid/Guided_Analytics_for_ML_Automation/01_Guided_Analytics_for_ML_Automation?exec=31a9e115-b00b-4f11-bfb2-93c218f9648f&single
Open for Innovation KNIME WebPortal
Administration
Logout
Open for Innovation KNIME
Upload Dataset
Select Target
Filter Columns
Select Models
Parameter Settings
Feature Eng. Settings
Execution Settings
Download Models

Select Models

Choose one or more machine learning models to train for your prediction task.

Simple models

☑ Naive Bayes

☑ Decision Tree

☑ Logistic Regression

Complex models

☐ Support Vector Machine

☑ Random Forest

☑ Generalized Linear Models

☑ Gradient Boosted Trees

☐ Deep Learning

Fine-tune Model Parameters

By default (unchecked), all parameters for the selected models and for feature engineering are automatically fully optimized. However, by checking this option you can guide the automatic optimization process.

☑ **Finetune Model Parameters**

Outlier Treatment

Guide

Select Models

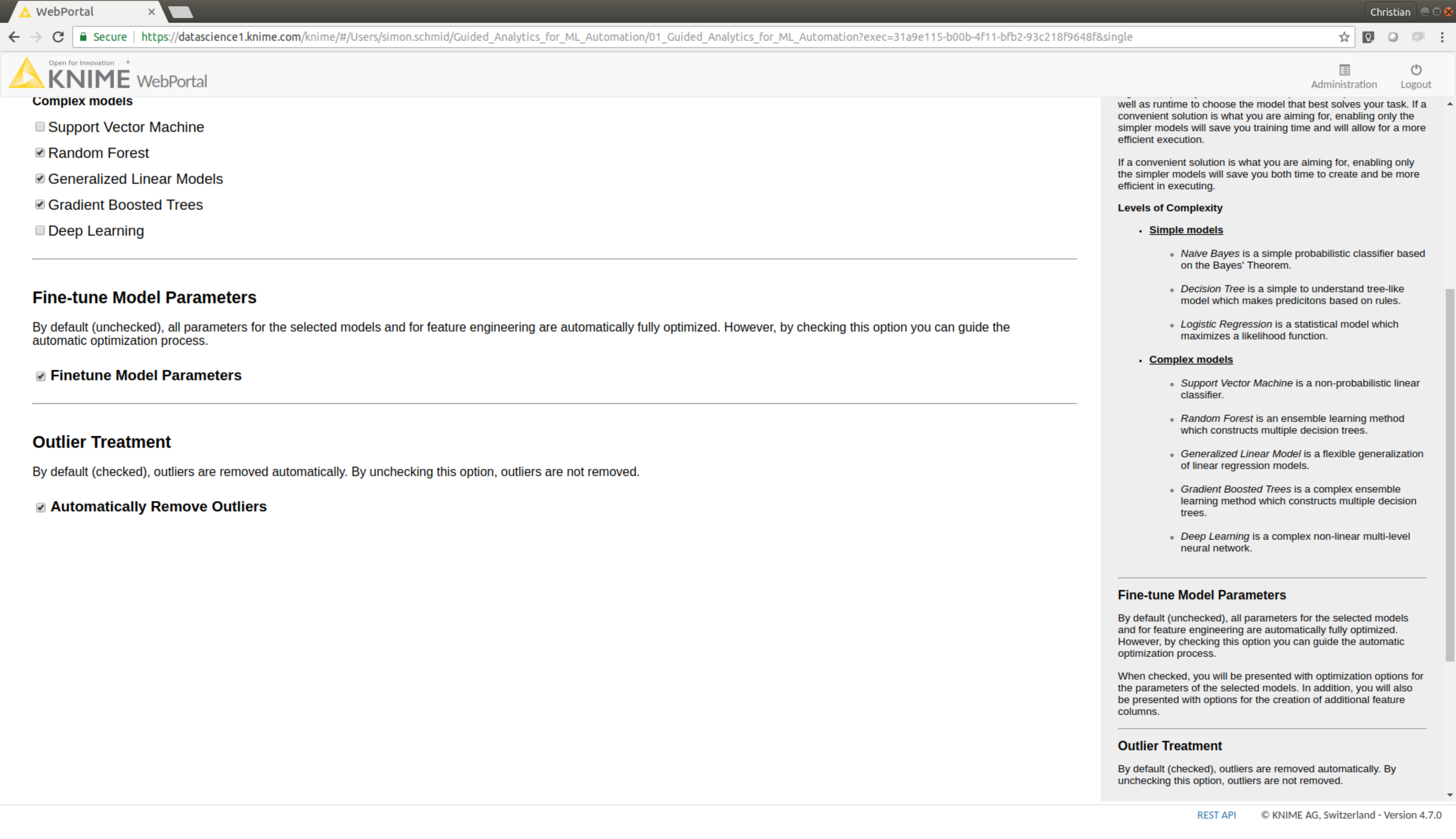
Choose which models you want to train. The available models have different levels of complexity. Less complex models are simpler to interpret and understand and generally faster to train and more efficient to use in production. In contrast, more complex models are capable of solving more complicated problems at possibly a finer level of detail but possibly at the cost of longer training times and less efficient usage in production. If the time is not an issue and you simply want to see the best performing model, use all of the proposed models, especially the ones with higher complexity. You can later compare model performances as well as runtime to choose the model that best solves your task. If a convenient solution is what you are aiming for, enabling only the simpler models will save you training time and will allow for a more efficient execution.

If a convenient solution is what you are aiming for, enabling only the simpler models will save you both time to create and be more efficient in executing.

Levels of Complexity

- Simple models
  - *Naive Bayes* is a simple probabilistic classifier based on the Bayes' Theorem.
  - *Decision Tree* is a simple to understand tree-like model which makes predicitons based on rules.
  - *Logistic Regression* is a statistical model which maximizes a likelihood function.
- Complex models
  - *Support Vector Machine* is a non-probabilistic linear classifier.
  - *Random Forest* is an ensemble learning method which constructs multiple decision trees.
  - *Generalized Linear Model* is a flexible generalization

© 2018 KNIME AG. All rights reserved.


WebPortal
Christian
Secure
https://datascience1.knime.com/knime/#/Users/simon.schmid/Guided_Analytics_for_ML_Automation/01_Guided_Analytics_for_ML_Automation?exec=31a9e115-b00b-4f11-bfb2-93c218f9648f&single
Open for Innovation
KNIME WebPortal
Administration
Logout

Complex models

☐ Support Vector Machine

☑ Random Forest

☑ Generalized Linear Models

☑ Gradient Boosted Trees

☐ Deep Learning

Fine-tune Model Parameters

By default (unchecked), all parameters for the selected models and for feature engineering are automatically fully optimized. However, by checking this option you can guide the automatic optimization process.

☑ **Finetune Model Parameters**

Outlier Treatment

By default (checked), outliers are removed automatically. By unchecking this option, outliers are not removed.

☑ **Automatically Remove Outliers**

well as runtime to choose the model that best solves your task. If a convenient solution is what you are aiming for, enabling only the simpler models will save you training time and will allow for a more efficient execution.

If a convenient solution is what you are aiming for, enabling only the simpler models will save you both time to create and be more efficient in executing.

Levels of Complexity

- **Simple models**
  - *Naive Bayes* is a simple probabilistic classifier based on the Bayes' Theorem.
  - *Decision Tree* is a simple to understand tree-like model which makes predicitons based on rules.
  - *Logistic Regression* is a statistical model which maximizes a likelihood function.
- **Complex models**
  - *Support Vector Machine* is a non-probabilistic linear classifier.
  - *Random Forest* is an ensemble learning method which constructs multiple decision trees.
  - *Generalized Linear Model* is a flexible generalization of linear regression models.
  - *Gradient Boosted Trees* is a complex ensemble learning method which constructs multiple decision trees.
  - *Deep Learning* is a complex non-linear multi-level neural network.

Fine-tune Model Parameters

By default (unchecked), all parameters for the selected models and for feature engineering are automatically fully optimized. However, by checking this option you can guide the automatic optimization process.

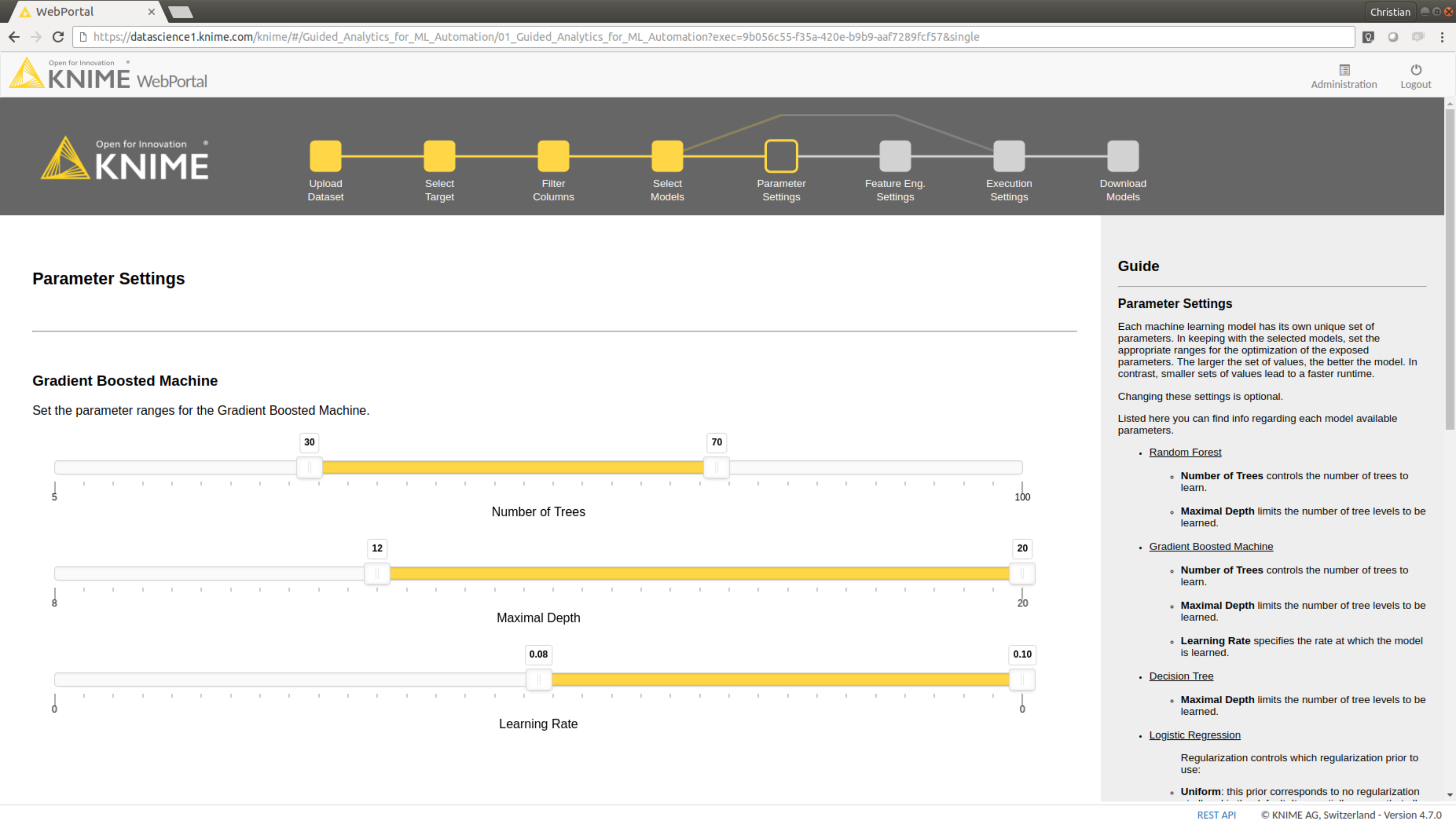
When checked, you will be presented with optimization options for the parameters of the selected models. In addition, you will also be presented with options for the creation of additional feature columns.

Outlier Treatment

By default (checked), outliers are removed automatically. By unchecking this option, outliers are not removed.

REST API © KNIME AG, Switzerland - Version 4.7.0

© 2018 KNIME AG. All rights reserved.


WebPortal
https://datascience1.knime.com/knime/#/Guided_Analytics_for_ML_Automation/01_Guided_Analytics_for_ML_Automation?exec=9b056c55-f35a-420e-b9b9-aaf7289fcf57&single
Christian
Open for Innovation KNIME WebPortal
Administration
Logout
Open for Innovation KNIME
Upload Dataset
Select Target
Filter Columns
Select Models
Parameter Settings
Feature Eng. Settings
Execution Settings
Download Models

Parameter Settings

Gradient Boosted Machine

Set the parameter ranges for the Gradient Boosted Machine.

Number of Trees: 30 – 70 (range 5 to 100)

Maximal Depth: 12 – 20 (range 8 to 20)

Learning Rate: 0.08 – 0.10 (range 0 to 0)

Guide

Parameter Settings

Each machine learning model has its own unique set of parameters. In keeping with the selected models, set the appropriate ranges for the optimization of the exposed parameters. The larger the set of values, the better the model. In contrast, smaller sets of values lead to a faster runtime.

Changing these settings is optional.

Listed here you can find info regarding each model available parameters.

- Random Forest
  - **Number of Trees** controls the number of trees to learn.
  - **Maximal Depth** limits the number of tree levels to be learned.
- Gradient Boosted Machine
  - **Number of Trees** controls the number of trees to learn.
  - **Maximal Depth** limits the number of tree levels to be learned.
  - **Learning Rate** specifies the rate at which the model is learned.
- Decision Tree
  - **Maximal Depth** limits the number of tree levels to be learned.
- Logistic Regression

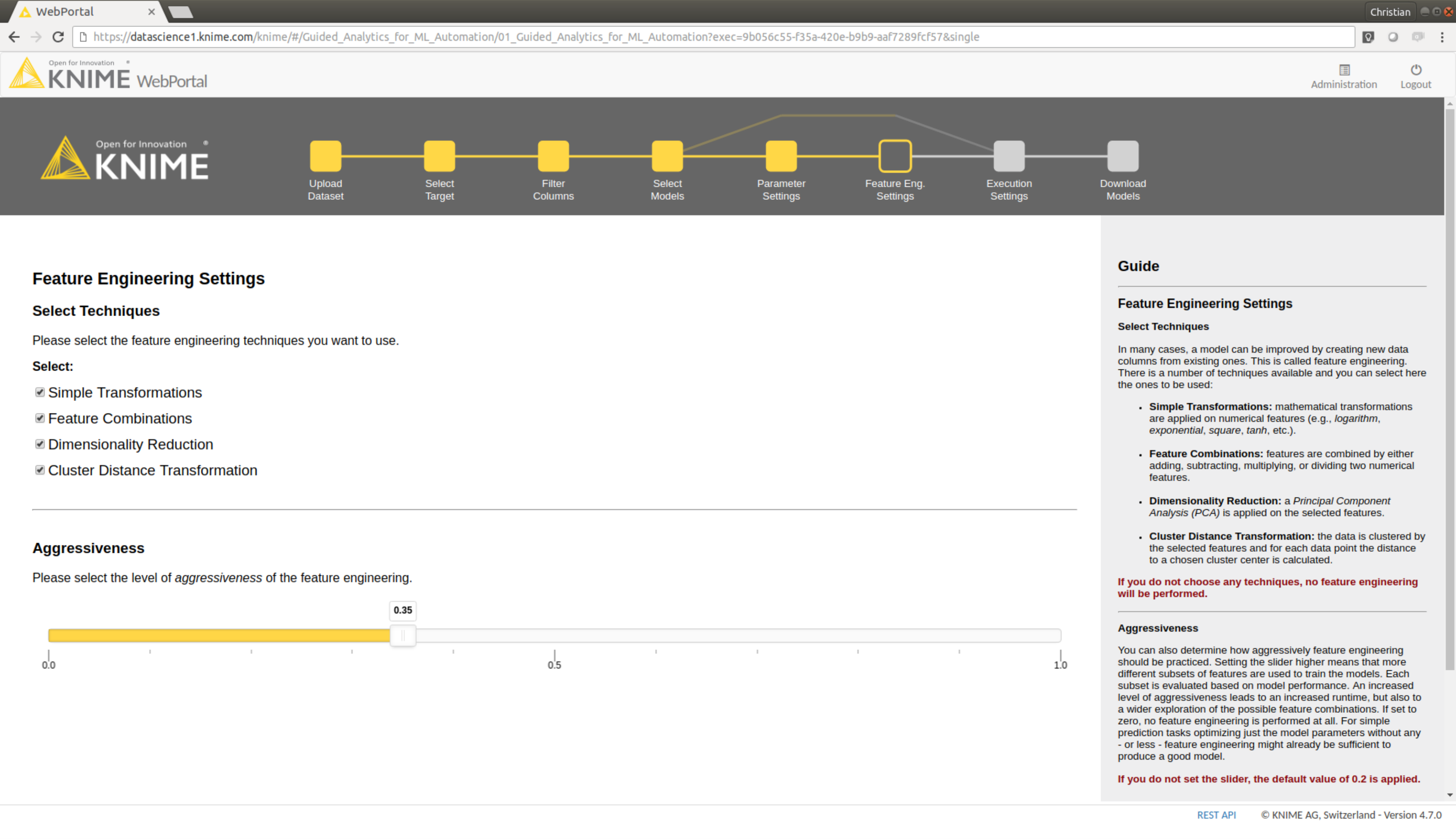
  Regularization controls which regularization prior to use:

  - **Uniform**: this prior corresponds to no regularization

REST API © KNIME AG, Switzerland - Version 4.7.0

© 2018 KNIME AG. All rights reserved.

Open for Innovation KNIME


WebPortal
https://datascience1.knime.com/knime/#/Guided_Analytics_for_ML_Automation/01_Guided_Analytics_for_ML_Automation?exec=9b056c55-f35a-420e-b9b9-aaf7289fcf57&single
Christian
Open for Innovation KNIME WebPortal
Administration
Logout
Open for Innovation KNIME
Upload Dataset
Select Target
Filter Columns
Select Models
Parameter Settings
Feature Eng. Settings
Execution Settings
Download Models

Feature Engineering Settings

Select Techniques

Please select the feature engineering techniques you want to use.

Select:

☑ Simple Transformations

☑ Feature Combinations

☑ Dimensionality Reduction

☑ Cluster Distance Transformation

Aggressiveness

Please select the level of *aggressiveness* of the feature engineering.

Guide

Feature Engineering Settings

Select Techniques

In many cases, a model can be improved by creating new data columns from existing ones. This is called feature engineering. There is a number of techniques available and you can select here the ones to be used:

- **Simple Transformations:** mathematical transformations are applied on numerical features (e.g., *logarithm*, *exponential*, *square*, *tanh*, etc.).
- **Feature Combinations:** features are combined by either adding, subtracting, multiplying, or dividing two numerical features.
- **Dimensionality Reduction:** a *Principal Component Analysis (PCA)* is applied on the selected features.
- **Cluster Distance Transformation:** the data is clustered by the selected features and for each data point the distance to a chosen cluster center is calculated.

If you do not choose any techniques, no feature engineering will be performed.

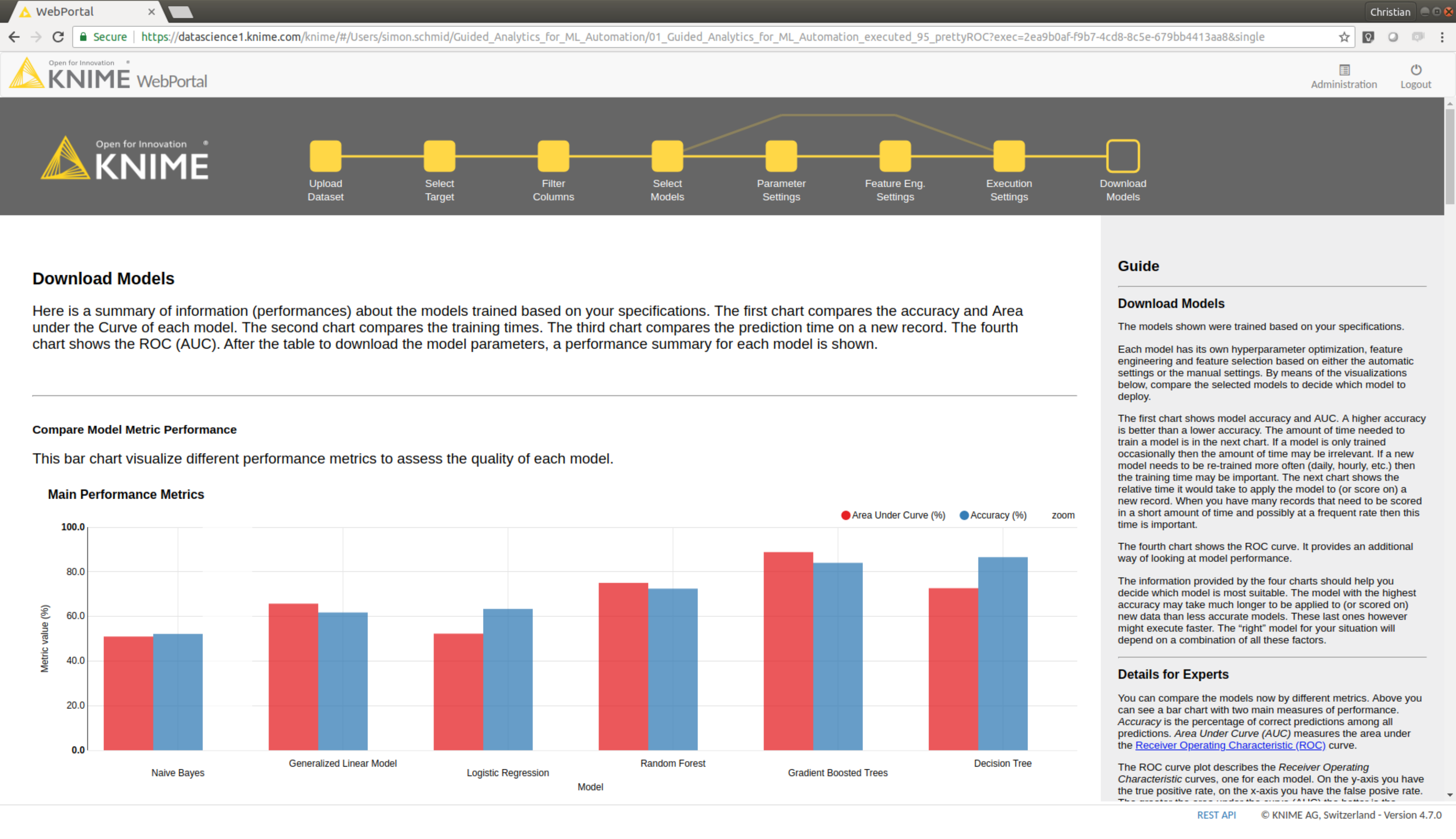
Aggressiveness

You can also determine how aggressively feature engineering should be practiced. Setting the slider higher means that more different subsets of features are used to train the models. Each subset is evaluated based on model performance. An increased level of aggressiveness leads to an increased runtime, but also to a wider exploration of the possible feature combinations. If set to zero, no feature engineering is performed at all. For simple prediction tasks optimizing just the model parameters without any - or less - feature engineering might already be sufficient to produce a good model.

If you do not set the slider, the default value of 0.2 is applied.

REST API © KNIME AG, Switzerland - Version 4.7.0

© 2018 KNIME AG. All rights reserved.



WebPortal
Christian
Secure
https://datascience1.knime.com/knime/#/Users/simon.schmid/Guided_Analytics_for_ML_Automation/01_Guided_Analytics_for_ML_Automation_executed_95_prettyROC?exec=2ea9b0af-f9b7-4cd8-8c5e-679bb4413aa8&single
Open for Innovation KNIME WebPortal
Administration
Logout
Upload Dataset
Select Target
Filter Columns
Select Models
Parameter Settings
Feature Eng. Settings
Execution Settings
Download Models

Download Models

Here is a summary of information (performances) about the models trained based on your specifications. The first chart compares the accuracy and Area under the Curve of each model. The second chart compares the training times. The third chart compares the prediction time on a new record. The fourth chart shows the ROC (AUC). After the table to download the model parameters, a performance summary for each model is shown.

Compare Model Metric Performance

This bar chart visualize different performance metrics to assess the quality of each model.

Main Performance Metrics

Guide

Download Models

The models shown were trained based on your specifications.

Each model has its own hyperparameter optimization, feature engineering and feature selection based on either the automatic settings or the manual settings. By means of the visualizations below, compare the selected models to decide which model to deploy.

The first chart shows model accuracy and AUC. A higher accuracy is better than a lower accuracy. The amount of time needed to train a model is in the next chart. If a model is only trained occasionally then the amount of time may be irrelevant. If a new model needs to be re-trained more often (daily, hourly, etc.) then the training time may be important. The next chart shows the relative time it would take to apply the model to (or score on) a new record. When you have many records that need to be scored in a short amount of time and possibly at a frequent rate then this time is important.

The fourth chart shows the ROC curve. It provides an additional way of looking at model performance.

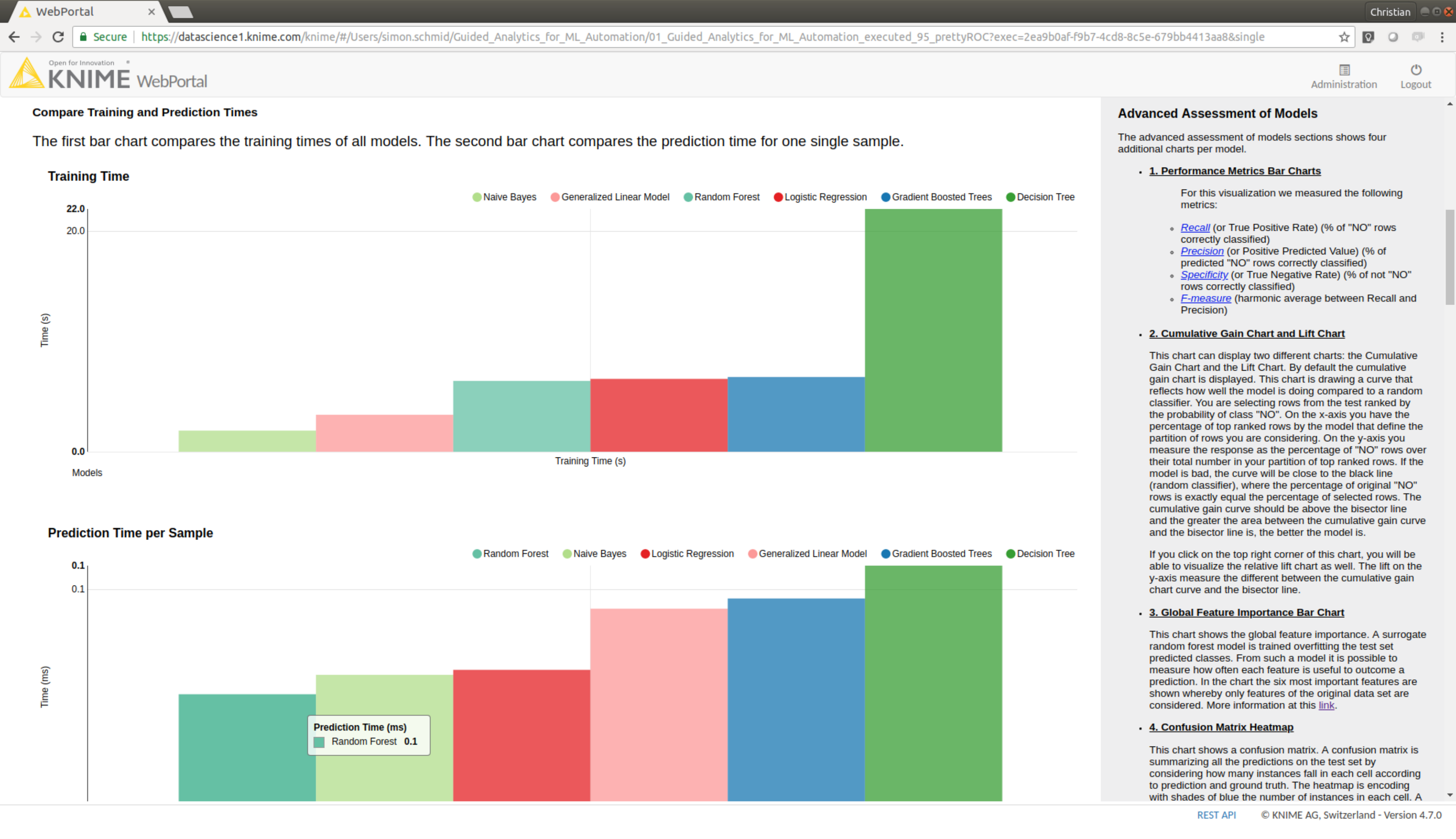
The information provided by the four charts should help you decide which model is most suitable. The model with the highest accuracy may take much longer to be applied to (or scored on) new data than less accurate models. These last ones however might execute faster. The “right” model for your situation will depend on a combination of all these factors.

Details for Experts

You can compare the models now by different metrics. Above you can see a bar chart with two main measures of performance. *Accuracy* is the percentage of correct predictions among all predictions. *Area Under Curve (AUC)* measures the area under the Receiver Operating Characteristic (ROC) curve.

The ROC curve plot describes the *Receiver Operating Characteristic* curves, one for each model. On the y-axis you have the true positive rate, on the x-axis you have the false posive rate. The greater the area under the curve (AUC) the better is the

© 2018 KNIME AG. All rights reserved.


WebPortal
Christian
Secure
https://datascience1.knime.com/knime/#/Users/simon.schmid/Guided_Analytics_for_ML_Automation/01_Guided_Analytics_for_ML_Automation_executed_95_prettyROC?exec=2ea9b0af-f9b7-4cd8-8c5e-679bb4413aa8&single
Open for Innovation
KNIME WebPortal
Administration
Logout

Compare Training and Prediction Times

The first bar chart compares the training times of all models. The second bar chart compares the prediction time for one single sample.

Training Time

Prediction Time per Sample

Advanced Assessment of Models

The advanced assessment of models sections shows four additional charts per model.

- **1. Performance Metrics Bar Charts**

  For this visualization we measured the following metrics:

  - *Recall* (or True Positive Rate) (% of "NO" rows correctly classified)
  - *Precision* (or Positive Predicted Value) (% of predicted "NO" rows correctly classified)
  - *Specificity* (or True Negative Rate) (% of not "NO" rows correctly classified)
  - *F-measure* (harmonic average between Recall and Precision)

- **2. Cumulative Gain Chart and Lift Chart**

  This chart can display two different charts: the Cumulative Gain Chart and the Lift Chart. By default the cumulative gain chart is displayed. This chart is drawing a curve that reflects how well the model is doing compared to a random classifier. You are selecting rows from the test ranked by the probability of class "NO". On the x-axis you have the percentage of top ranked rows by the model that define the partition of rows you are considering. On the y-axis you measure the response as the percentage of "NO" rows over their total number in your partition of top ranked rows. If the model is bad, the curve will be close to the black line (random classifier), where the percentage of original "NO" rows is exactly equal the percentage of selected rows. The cumulative gain curve should be above the bisector line and the greater the area between the cumulative gain curve and the bisector line is, the better the model is.

  If you click on the top right corner of this chart, you will be able to visualize the relative lift chart as well. The lift on the y-axis measure the different between the cumulative gain chart curve and the bisector line.

- **3. Global Feature Importance Bar Chart**

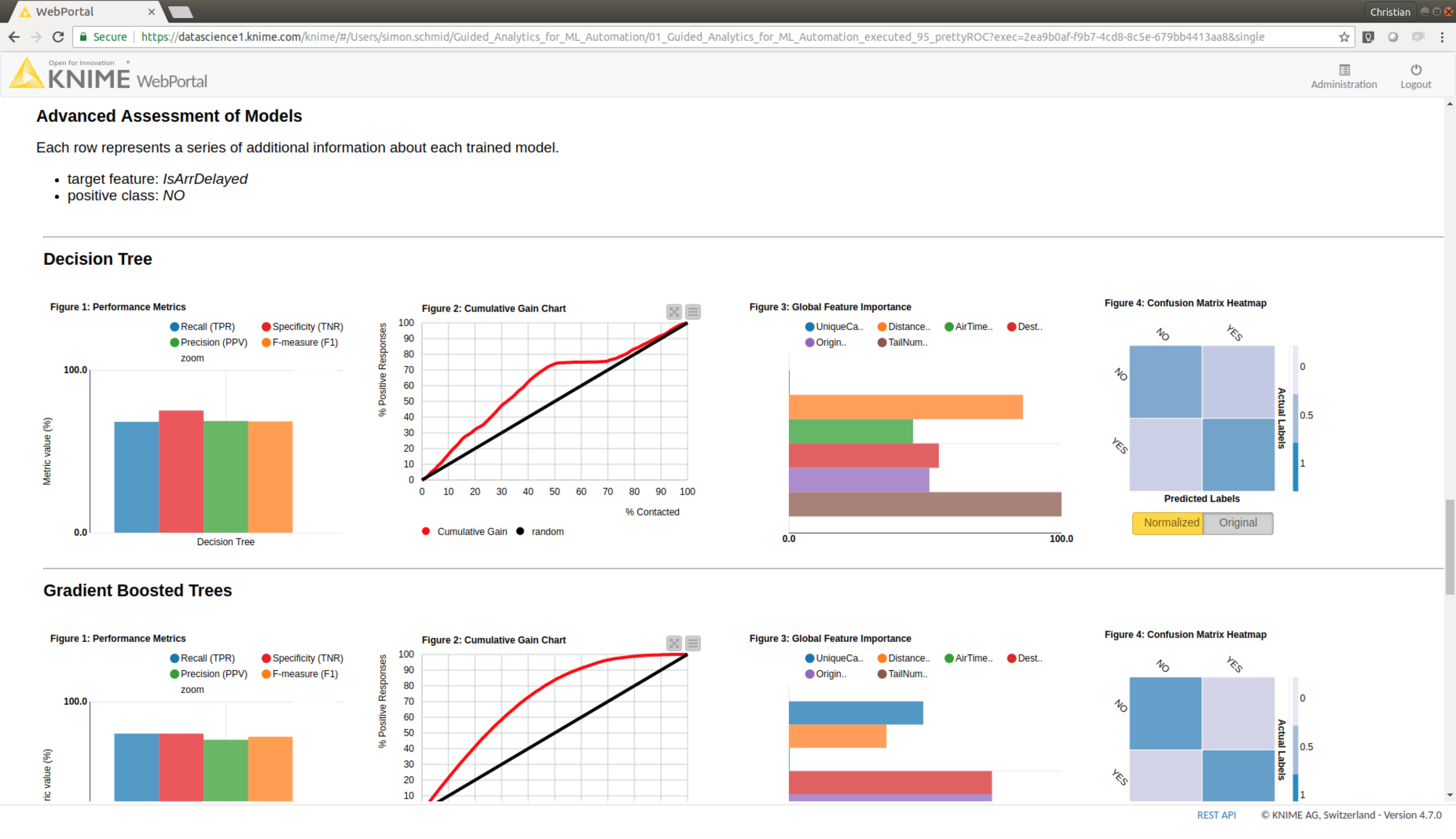
  This chart shows the global feature importance. A surrogate random forest model is trained overfitting the test set predicted classes. From such a model it is possible to measure how often each feature is useful to outcome a prediction. In the chart the six most important features are shown whereby only features of the original data set are considered. More information at this link.

- **4. Confusion Matrix Heatmap**

  This chart shows a confusion matrix. A confusion matrix is summarizing all the predictions on the test set by considering how many instances fall in each cell according to prediction and ground truth. The heatmap is encoding with shades of blue the number of instances in each cell. A

REST API © KNIME AG, Switzerland - Version 4.7.0

© 2018 KNIME AG. All rights reserved.



WebPortal
Christian
Secure
https://datascience1.knime.com/knime/#/Users/simon.schmid/Guided_Analytics_for_ML_Automation/01_Guided_Analytics_for_ML_Automation_executed_95_prettyROC?exec=2ea9b0af-f9b7-4cd8-8c5e-679bb4413aa8&single
Open for Innovation
KNIME WebPortal
Administration
Logout

Advanced Assessment of Models

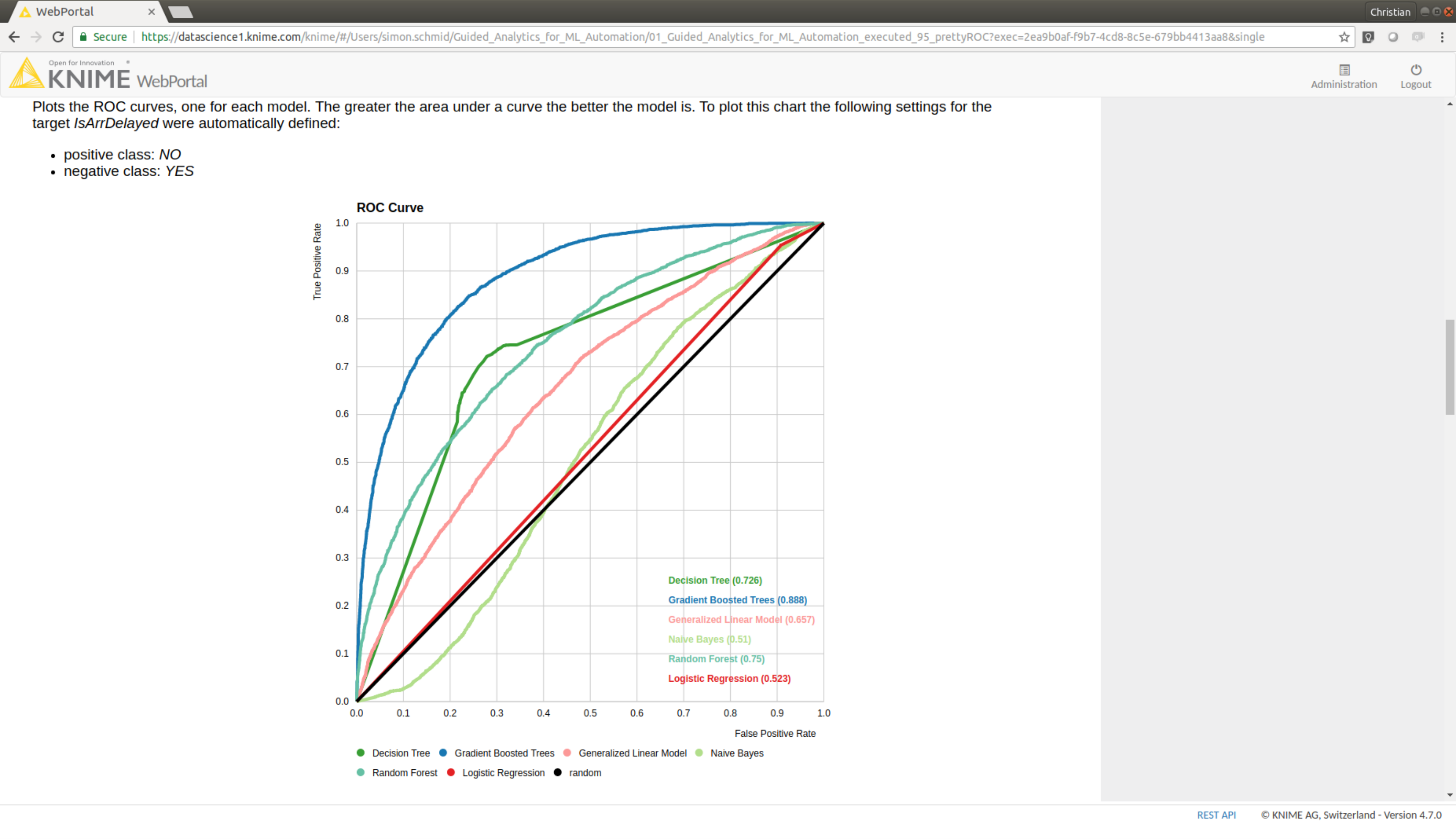
Each row represents a series of additional information about each trained model.

- target feature: *IsArrDelayed*
- positive class: *NO*

Decision Tree

Gradient Boosted Trees

© 2018 KNIME AG. All rights reserved.



WebPortal
Christian
Secure
https://datascience1.knime.com/knime/#/Users/simon.schmid/Guided_Analytics_for_ML_Automation/01_Guided_Analytics_for_ML_Automation_executed_95_prettyROC?exec=2ea9b0af-f9b7-4cd8-8c5e-679bb4413aa8&single
Open for Innovation
KNIME WebPortal
Administration
Logout

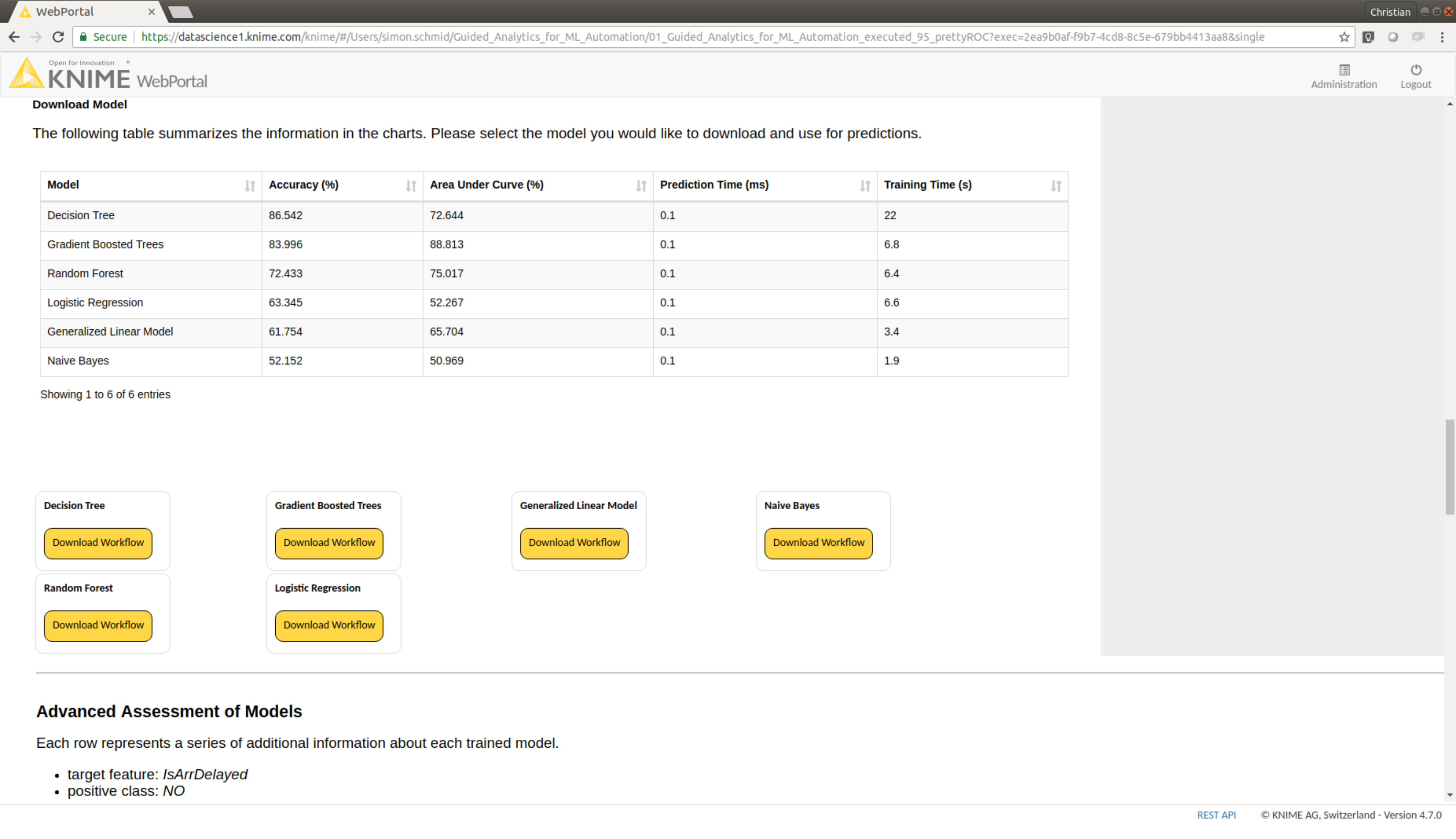
Plots the ROC curves, one for each model. The greater the area under a curve the better the model is. To plot this chart the following settings for the target *IsArrDelayed* were automatically defined:

- positive class: *NO*
- negative class: *YES*

ROC Curve
True Positive Rate
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
False Positive Rate
Decision Tree (0.726)
Gradient Boosted Trees (0.888)
Generalized Linear Model (0.657)
Naive Bayes (0.51)
Random Forest (0.75)
Logistic Regression (0.523)
● Decision Tree
● Gradient Boosted Trees
● Generalized Linear Model
● Naive Bayes
● Random Forest
● Logistic Regression
● random

REST API
© KNIME AG, Switzerland - Version 4.7.0

© 2018 KNIME AG. All rights reserved.


WebPortal
Christian
Secure
https://datascience1.knime.com/knime/#/Users/simon.schmid/Guided_Analytics_for_ML_Automation/01_Guided_Analytics_for_ML_Automation_executed_95_prettyROC?exec=2ea9b0af-f9b7-4cd8-8c5e-679bb4413aa8&single
Open for Innovation
KNIME WebPortal
Administration
Logout

Download Model

The following table summarizes the information in the charts. Please select the model you would like to download and use for predictions.

Model	Accuracy (%)	Area Under Curve (%)	Prediction Time (ms)	Training Time (s)
Decision Tree	86.542	72.644	0.1	22
Gradient Boosted Trees	83.996	88.813	0.1	6.8
Random Forest	72.433	75.017	0.1	6.4
Logistic Regression	63.345	52.267	0.1	6.6
Generalized Linear Model	61.754	65.704	0.1	3.4
Naive Bayes	52.152	50.969	0.1	1.9

Showing 1 to 6 of 6 entries

Decision Tree

Gradient Boosted Trees

Generalized Linear Model

Naive Bayes

Random Forest

Logistic Regression

Advanced Assessment of Models

Each row represents a series of additional information about each trained model.

- target feature: *IsArrDelayed*
- positive class: *NO*

© KNIME AG, Switzerland - Version 4.7.0

© 2018 KNIME AG. All rights reserved.

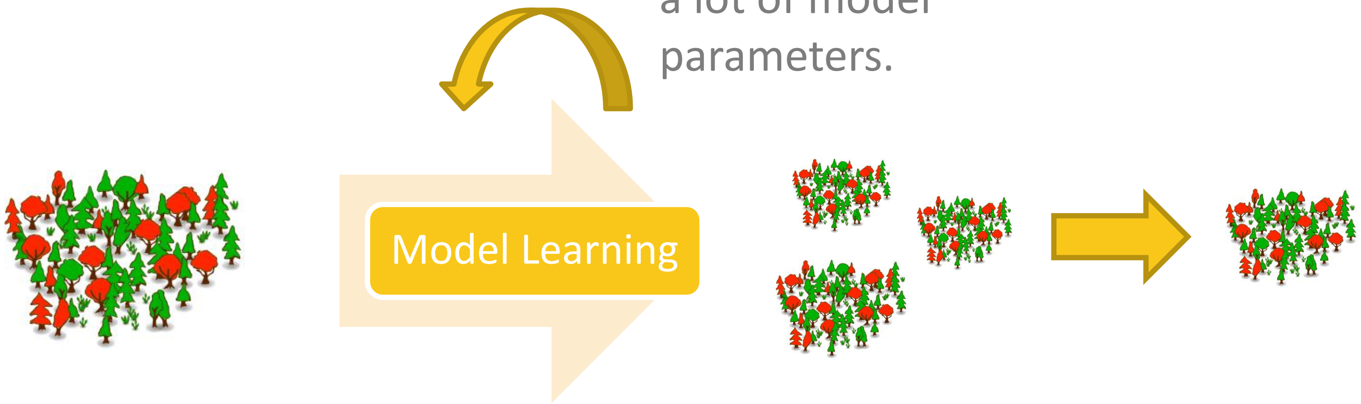
What do we Automate?

- Data Cleansing
  - Missing value handling, calculate statistics, outlier detection
- Feature Engineering
  - Mathematical transformations, feature combinations, dimensionality reduction and more
- Feature Selection
  - Forward feature selection, backward feature elimination, genetic algorithm etc.

© 2018 KNIME AG. All rights reserved.

What do we Automate?

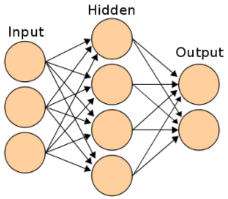
- Parameter Optimization

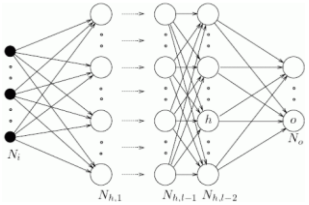


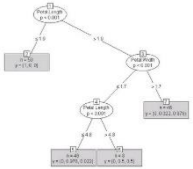
Automatically try a lot of model parameters.

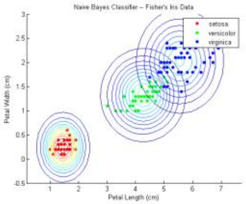
What do we Automate?

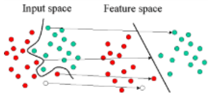
- Model Selection
  - Try many models, but in an automated way.







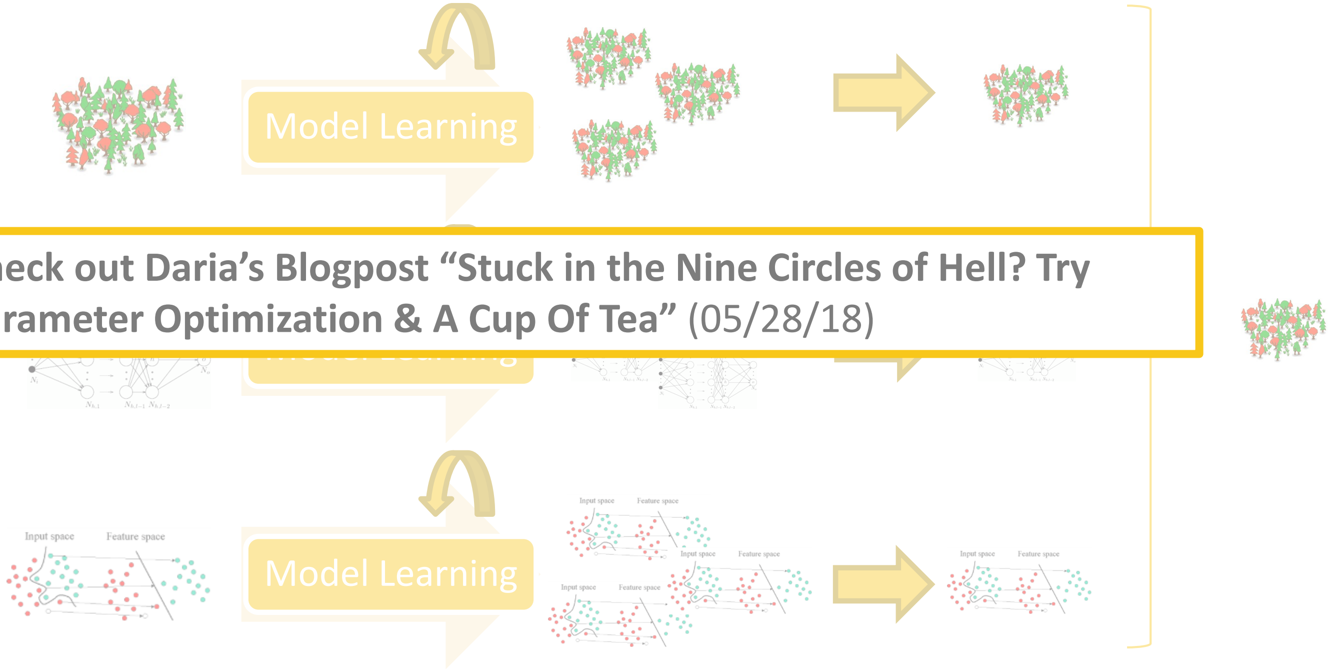




© 2018 KNIME AG. All rights reserved.

What do we Automate?

- Model Selection and Parametrization



Check out Daria's Blogpost "Stuck in the Nine Circles of Hell? Try Parameter Optimization & A Cup Of Tea" (05/28/18)

© 2018 KNIME AG. All rights reserved.