Guided Analytics for Machine Learning Automation

Christian Dietz, Simon Schmid
KNIME
Automating Everything?

Automating Data Integration:
- Parsing
- Record Matching
- ...

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

Automating Data Proc:
- Feature Selection
- Feature Construction
- Data Cleaning

Automating Analytics:
- Parameter Optimization
- Model Selection
- Ensemble Construction
Automating Everything?

Automating Data Proc:
- Feature Selection
- Feature Construction
- Data Cleaning

Automating Analytics:
- Parameter Optimization
- Model Selection
- Ensemble Construction

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

Automating Data Integration:
- Parsing
- Record Matching
- ...
Building a Guided Automation Workflow

Interaction Points

Automated
Guided Automation: Automation + Interaction
Guided Automation: Automation + Interaction

Display column information and allow the user to exclude columns.

Data Explorer (JavaScript)
Exclude manually feature columns.

Table View (JavaScript)
Display column metrics.

Range Slider
Color Manager
Filter Definition
Filter Apply

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. To help you decide, the overall column relevance towards the final prediction is measured:

- **Column Relevance** measures overall metric, correlating the metrics below. Use the slider to select the most relevant columns based on their Column Relevance.

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

- **ID/Noise Test** measures how likely the column is a random variable. The lower the ID/Noise Test, the less likely the column is random. If these values are problematic for your project, it should be removed.

- **Constant Value Test** measures how likely the column contains a set of constant values. Columns with a constant value should not have information. You should avoid using them.

- **Missing Value Test** measures the percentage of missing values in a column over the entire dataset. You should consider removing columns with a high percentage of missing values.

By using the slider, columns can be evaluated from model training toward their relevance.

Furthermore, you can use the linear correlation between each column and the outcome to predict your output values.

- **Correlation with Target** measures how well the column will predict the outcome. It is important to keep or remove a column based on the correlation. If you have high correlation (close to 1 or -1) this will help the model.
Guided Automation: Automation + Interaction
Guided Automation: Automation + Interaction
Guided Automation: Automation + Interaction

WebPortal

Workflow

KNIME Server
Guided Automation on KNIME Server

Live Demo
Scoring Workflow
Scoring Workflow

KNIME Server
Customize the Blueprint for Text Processing

Added by in-house expert.
Thank You!

Download workflow from knime.com and get started!
Upload Dataset

Upload the dataset to be used.

Selected file "airline.csv" (4 MB)
### Select Target

Select the target column whose values should be predicted.

**Select:**

<table>
<thead>
<tr>
<th>Row ID</th>
<th>Month</th>
<th>DayOfWeek</th>
<th>UniqueCarrier</th>
<th>Cancelled</th>
<th>CancellationCode</th>
<th>Diverted</th>
<th>WeatherDelay</th>
<th>SecurityDelay</th>
<th>IsArrDelayed</th>
<th>IsDepDelayed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row0</td>
<td>10</td>
<td>3</td>
<td>PS</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>?</td>
<td>?</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Row1</td>
<td>10</td>
<td>4</td>
<td>PS</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>?</td>
<td>?</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Row2</td>
<td>10</td>
<td>6</td>
<td>PS</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>?</td>
<td>?</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>Row3</td>
<td>10</td>
<td>7</td>
<td>PS</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>?</td>
<td>?</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>Row4</td>
<td>10</td>
<td>1</td>
<td>PS</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>?</td>
<td>?</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Row5</td>
<td>10</td>
<td>3</td>
<td>PS</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>?</td>
<td>?</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>Row6</td>
<td>10</td>
<td>4</td>
<td>PS</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>?</td>
<td>?</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Row7</td>
<td>10</td>
<td>5</td>
<td>PS</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>?</td>
<td>?</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Row8</td>
<td>10</td>
<td>6</td>
<td>PS</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>?</td>
<td>?</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Row9</td>
<td>10</td>
<td>7</td>
<td>PS</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>?</td>
<td>?</td>
<td>YES</td>
<td>NO</td>
</tr>
</tbody>
</table>
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.

Overall Column Relevance

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Overall Column Relevance</th>
<th>Correlation with Target (%)</th>
<th>ID/Noise Test (%)</th>
<th>Constant Value Test (%)</th>
<th>Missing Value Test (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRSElapsedTime</td>
<td>96.44</td>
<td>28.756</td>
<td>3.04</td>
<td>3.56</td>
<td>0.04</td>
</tr>
<tr>
<td>ArrDelay</td>
<td>96.39</td>
<td>54.009</td>
<td>2.42</td>
<td>3.61</td>
<td>2.69</td>
</tr>
<tr>
<td>ActualElapsedTime</td>
<td>96.3</td>
<td>26.456</td>
<td>3.7</td>
<td>2.69</td>
<td>2.69</td>
</tr>
<tr>
<td>CRSDepTime</td>
<td>95.19</td>
<td>18.621</td>
<td>4.81</td>
<td>2.49</td>
<td>0</td>
</tr>
<tr>
<td>Year</td>
<td>94.96</td>
<td>23.379</td>
<td>0.21</td>
<td>3.04</td>
<td>0</td>
</tr>
<tr>
<td>Distance</td>
<td>94.239</td>
<td>18.965</td>
<td>5.761</td>
<td>4.69</td>
<td>0.07</td>
</tr>
<tr>
<td>DayofMonth</td>
<td>92.74</td>
<td>21.197</td>
<td>0.3</td>
<td>7.26</td>
<td>0</td>
</tr>
<tr>
<td>Origin</td>
<td>91.66</td>
<td>26.04</td>
<td>1.27</td>
<td>8.34</td>
<td>0</td>
</tr>
<tr>
<td>CO2AirTime</td>
<td>90.74</td>
<td>10.464</td>
<td>0.231</td>
<td>1.38</td>
<td>0</td>
</tr>
</tbody>
</table>
## Manually Select Columns

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

<table>
<thead>
<tr>
<th>Column</th>
<th>Exclude Column</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Median</th>
<th>Standard Deviation</th>
<th>Variance</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Overall Sum</th>
<th>No. zeros</th>
<th>No. missings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td></td>
<td>1987</td>
<td>2008</td>
<td>1997</td>
<td>1997</td>
<td>6.316</td>
<td>39.894</td>
<td>0.007</td>
<td>-1.193</td>
<td>19974554</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Month</td>
<td></td>
<td>1</td>
<td>10</td>
<td>1.399</td>
<td>1</td>
<td>1.852</td>
<td>3.430</td>
<td>4.430</td>
<td>17.629</td>
<td>1397</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DaysOfMonth</td>
<td></td>
<td>1</td>
<td>31</td>
<td>14.685</td>
<td>14</td>
<td>9.202</td>
<td>84.673</td>
<td>0.167</td>
<td>-1.222</td>
<td>148484</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DayOfWeek</td>
<td></td>
<td>1</td>
<td>7</td>
<td>3.843</td>
<td>4</td>
<td>1.910</td>
<td>3.648</td>
<td>0.154</td>
<td>-1.092</td>
<td>38433</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DepTime</td>
<td></td>
<td>1</td>
<td>2400</td>
<td>1345.779</td>
<td>1330</td>
<td>466.991</td>
<td>218080.269</td>
<td>0.085</td>
<td>-1.117</td>
<td>13129419</td>
<td>0</td>
<td>244</td>
</tr>
</tbody>
</table>

**Histogram**

- No. NaN: 0
- No. ==: 0
- No. >=: 0

[Graph showing histogram of data distribution]
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.

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 lower 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, disabling 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 Bayes’ Theorem.
  - Decision Tree is a simple and easy Decision Tree model which makes predictions 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
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.

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

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 predictions 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.
Parameter Settings

Gradient Boosted Machine

Set the parameter ranges for the Gradient Boosted Machine.

- Number of Trees
  - Range: 5 to 100
  - Default: 70
- Maximal Depth
  - Range: 1 to 20
  - Default: 29
- Learning Rate
  - Range: 0.01 to 1.0
  - Default: 0.08

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.
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.

0.35

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.
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

- Area Under Curve (%)
- Accuracy (%)

- Naive Bayes
- Generalized Linear Model
- Logistic Regression
- Random Forest
- Gradient Boosted Trees
- Decision Tree

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 rare 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 (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. Those 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 positive rate.
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**

- Naïve Bayes
- Generalized Linear Model
- Random Forest
- Logistic Regression
- Gradient Boosted Trees
- Decision Tree

**Prediction Time per Sample**

- Random Forest
- Naïve Bayes
- Logistic Regression
- Generalized Linear Model
- Gradient Boosted Trees
- Decision Tree

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), whereas 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 difference 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 out-of-the-box 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 on 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 correlating how many instances fall in each cell according to prediction and ground truth. The heatmap is encoded with shades of blue the number of instances in each cell. A
Advanced Assessment of Models

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

- target feature: IsArtDelayed
- positive class: NO

Decision Tree

Gradient Boosted Trees
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 IsAirDelayed were automatically defined:

- positive class: NO
- negative class: YES

![ROC Curve Diagram]

- Decision Tree (0.726)
- Gradient Boosted Trees (0.888)
- Generalized Linear Model (0.687)
- Naive Bayes (0.51)
- Random Forest (0.75)
- Logistic Regression (0.923)
The following table summarizes the information in the charts. Please select the model you would like to download and use for predictions.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
<th>Area Under Curve (%)</th>
<th>Prediction Time (ms)</th>
<th>Training Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree</td>
<td>89.542</td>
<td>72.644</td>
<td>0.1</td>
<td>22</td>
</tr>
<tr>
<td>Gradient Boosted Trees</td>
<td>83.996</td>
<td>88.813</td>
<td>0.1</td>
<td>6.8</td>
</tr>
<tr>
<td>Random Forest</td>
<td>72.433</td>
<td>75.017</td>
<td>0.1</td>
<td>6.4</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>63.345</td>
<td>52.267</td>
<td>0.1</td>
<td>6.6</td>
</tr>
<tr>
<td>Generalized Linear Model</td>
<td>61.754</td>
<td>65.704</td>
<td>0.1</td>
<td>3.4</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>52.152</td>
<td>50.069</td>
<td>0.1</td>
<td>1.9</td>
</tr>
</tbody>
</table>

Showing 1 to 6 of 6 entries

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

- target feature: IsArtDelayed
- positive class: NO
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.
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.
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)