

Lecture 8

Modeling with Linear Regression

Kim, Yang Sok

Dept. of MIS, Keimyung University

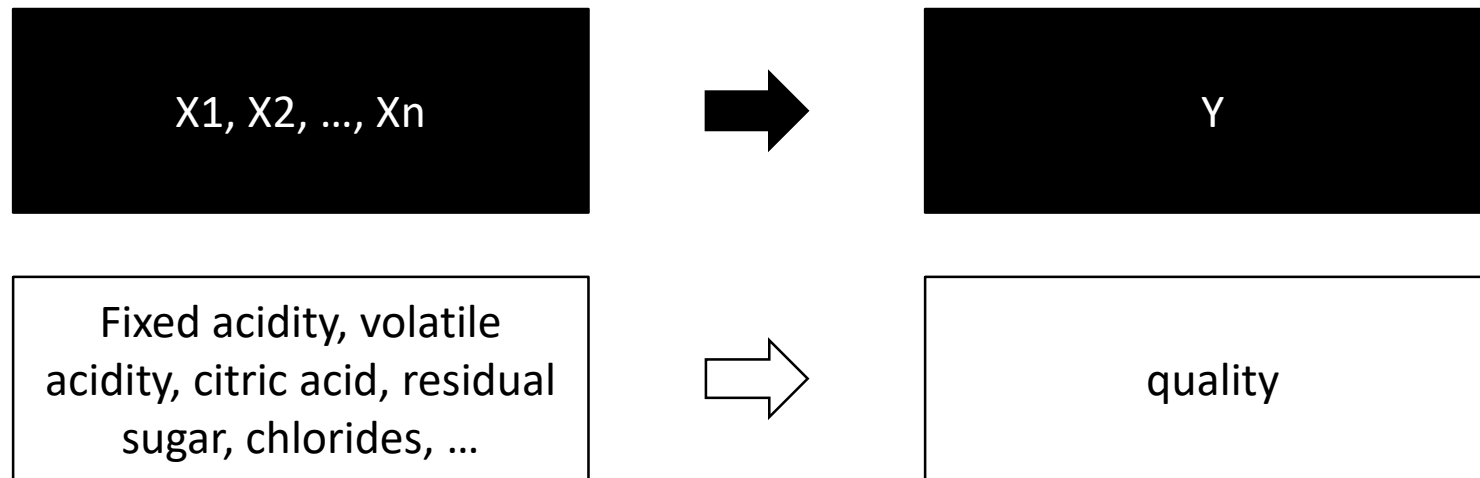
- **Introduction**
- **Linear Regression**
- **Exercises**
 - Exercise 1: Linear Regression with Rapidminer
 - Exercise 2: Linear Regression after Removing Insignificant Attributes
 - Exercise 3: Linear Regression with Variance Filter
 - Exercise 4: Linear Regression with Correlation Filter
 - Exercise 5: Linear Regression with Weight Filter
 - Exercise 6: Linear Regression with Forward Selection
 - Exercise 7: Linear Regression with Backward Elimination
 - Exercise 8: Linear Regression with Model Explainer
 - Exercise 9: Linear Regression with Model Simulator
- **Conclusion**

Introduction

- In this lecture, we also learn linear regression technique, a well-known prediction technique, accompanying with above issues
- In addition, we will focus on the following two issues
 - **Feature selection**: How to select **subset of attributes** that are appropriate for model learning?
 - **Parameter settings**: How to set appropriate parameters for model learning?

Linear Regression

- Linear regression is a method for modeling the relationship between one or more independent variables and a dependent variable.



fixed acidity	volatile acidity	citric acid	residual sug...	chlorides	free sulfur d...	total sulfur d...	density	pH	sulphates	alcohol	quality
7.400	0.700	0	1.900	0.076	11	34	0.998	3.510	0.560	9.400	5
7.800	0.880	0	2.600	0.098	25	67	0.997	3.200	0.680	9.800	5
7.800	0.760	0.040	2.300	0.092	15	54	0.997	3.260	0.650	9.800	5
11.200	0.280	0.560	1.900	0.075	17	60	0.998	3.160	0.580	9.800	6
7.400	0.700	0	1.900	0.076	11	34	0.998	3.510	0.560	9.400	5

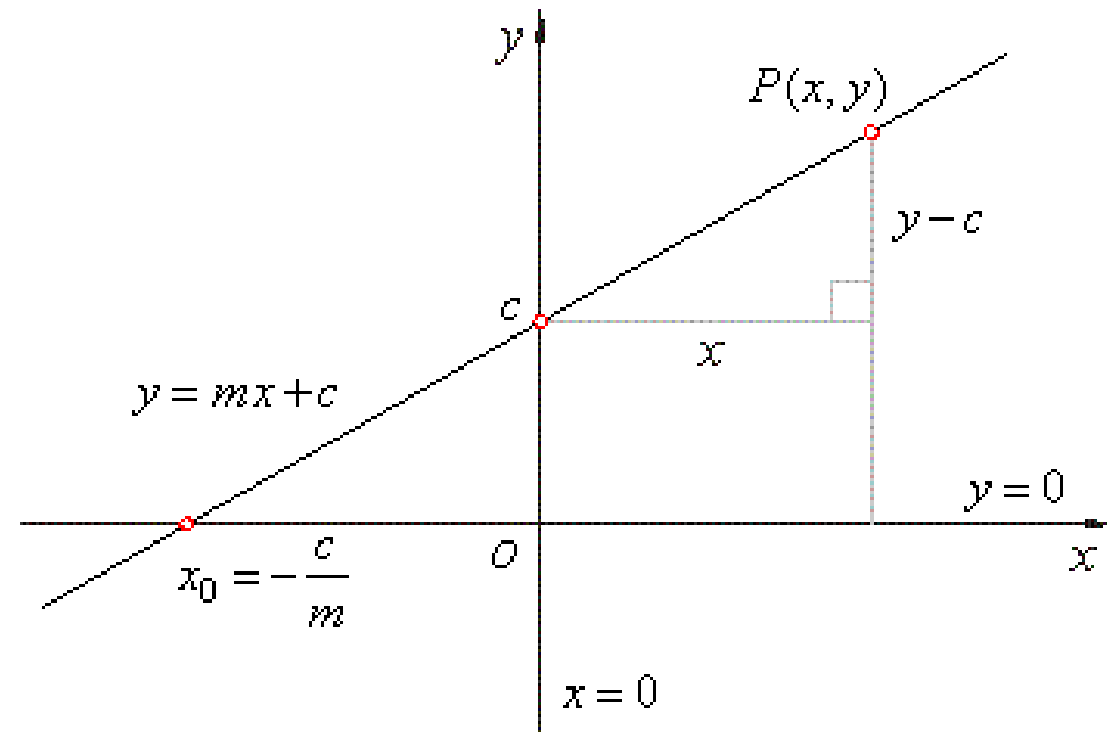
- Let X be the independent variable and Y be the dependent variable. We will define a linear relationship between these two variables as follows:

$$Y = mX + c$$

Intercept

Slope

How to
determine m
and c ?



- Linear regression was developed in the field of statistics, but has been borrowed by machine learning.
- Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).
- When there is a single input variable (x), the method is referred to as **simple linear regression**. When there are multiple input variables, literature from statistics often refers to the method as **multiple linear regression**.

- The loss is the error in our predicted value of m and c . **Our goal is to minimize this error to obtain the most accurate value of m and c .**
- We will use the **Mean Squared Error** function to calculate the loss. There are three steps in this function:
 1. Find the difference between the actual y and predicted y value($y = mx + c$), for a given x .
 2. Square this difference.
 3. Find the mean of the squares for every value in X .

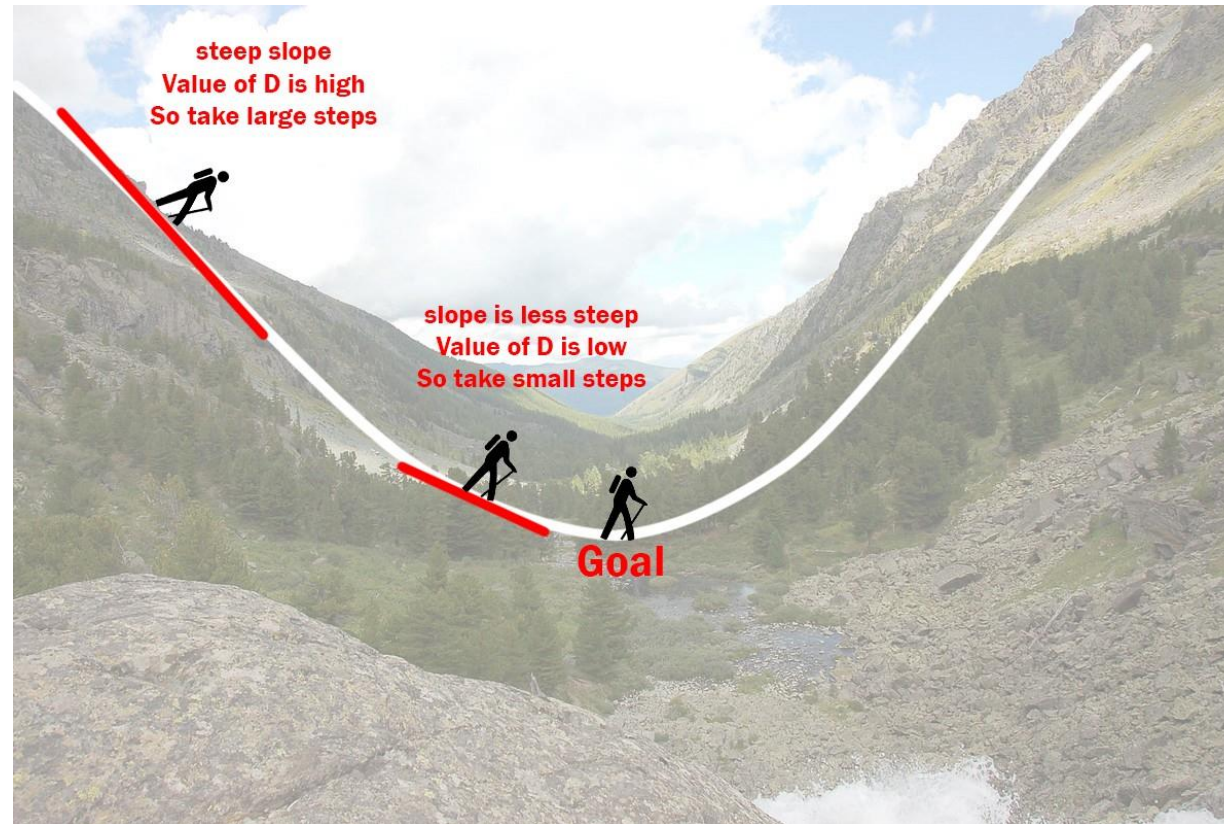
$$E = \frac{1}{n} \sum_{i=0}^n (y_i - \bar{y}_i)^2$$

- Here y_i is the actual value and \bar{y}_i is the predicted value. Lets substitute the value of \bar{y}_i :

$$E = \frac{1}{n} \sum_{i=0}^n (y_i - (mx_i + c))^2$$

- So we square the error and find the mean. hence the name Mean Squared Error. Now that we have defined the loss function, lets get into the interesting part — minimizing it and finding m and c .

- Gradient descent is an iterative optimization algorithm to find the minimum of a function. Here that function is our Loss Function.
- Understanding Gradient Descent



1. Initially let $m = 0$ and $c = 0$. Let L be our **learning rate**. This controls **how much the value of m changes with each step**. L could be a small value like 0.0001 for good accuracy.
2. Calculate **the partial derivative of the loss function** with respect to m , and plug in the current values of x , y , m and c in it to obtain the derivative value D

$$D_m = \frac{1}{n} \sum_{i=0}^n 2(y_i - (mx_i + c))(-x_i)$$
$$D_m = \frac{-2}{n} \sum_{i=0}^n x_i(y_i - \bar{y}_i)$$

D_m is the value of the partial derivative with respect to m . Similarly let's find the partial derivative with respect to c , D_c :

$$D_c = \frac{-2}{n} \sum_{i=0}^n (y_i - \bar{y}_i)$$

3. Now we update the current value of m and c using the following equation:

$$m = m - L \times D_m$$

$$c = c - L \times D_c$$

4. We repeat this process until our loss function is a very small value or ideally 0 (which means 0 error or 100% accuracy). The value of m and c that we are left with now will be the optimum values.

- **Linear Assumption**
 - Linear regression assumes that the **relationship between your input and output is linear**. This may be obvious, but it is good to remember when you have a lot of attributes. You may need to transform data to make the relationship linear (e.g. log transform for an exponential relationship).
- **Remove Noise**
 - Linear regression assumes that your input and output variables are **not noisy**. Consider using data cleaning operations that let you better expose and clarify the signal in your data. This is most important for the output variable and you want to remove outliers in the output variable (y) if possible.
- **Remove Collinearity**
 - Linear regression will over-fit your data when you have highly correlated input variables. Consider calculating pairwise correlations for your input data and removing the most correlated.

- **Gaussian Distributions**

- Linear regression will make more reliable predictions if your input and output variables have a Gaussian distribution. You may get some benefit using transforms (e.g. log or BoxCox) on your variables to make their distribution more Gaussian looking.

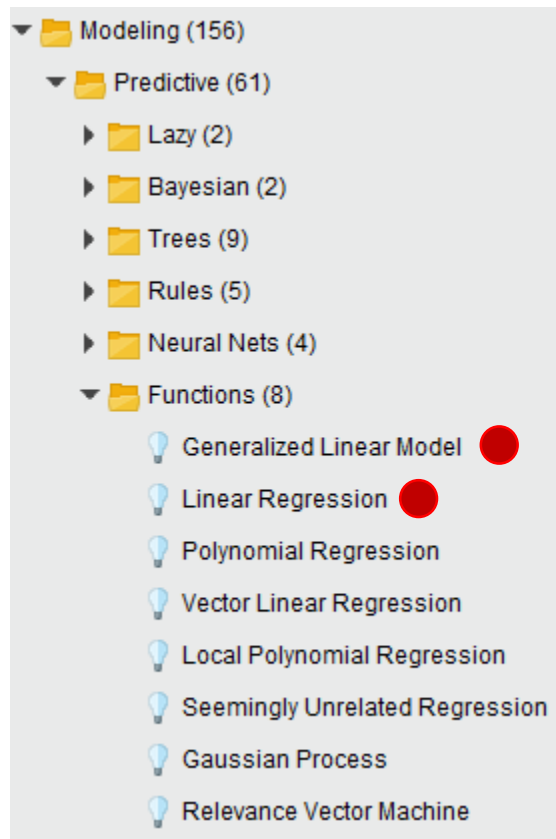
- **Rescale Inputs**

- Linear regression will often make more reliable predictions if you rescale input variables using **standardization or normalization**.

Linear Regression

Linear Regression in Rapidminer

- Rapidminer provides <<Linear Regression>> in 'Modeling>Predictive>Function' package



Exercise 1: Linear Regression with Rapidminer

Exercise 1: Linear Regression with Rapidminer

Task & Process

- **Task**
 - After loading data, build a linear regression model using **split test design**
- **Process**
 - Load “red wine” dataset
 - Set “quality” as label
 - Create split test design with <<Split Data>>
 - Create a linear regression model with the train dataset using linear regression algorithm
 - Apply the model to the test dataset
 - Set **regression** performance measures
 - Run the analysis process and check the performance results

Create split test design with <<Split Data>>

Views: Design Results Turbo Prep Auto Model

Find data, operators...etc All Studio

Repository

Operators

Split Data

Blending (1)

Examples (1)

Sampling (1)

Split Data

Extensions (1)

Statistics Extension (1)

Tools (1)

Split Data (by groups)

No results were found.

Recommended Operators

Business Data Mining ©Kim Yang Sok

Load "red wine" dataset

Split dataset into training and test dataset (7:3)

Set 'quality' as 'label'

Parameters

Process

logverbosity init

logfile

resultfile

random seed 2001

send mail never

encoding SYSTEM

Hide advanced parameters

Change compatibility (9.3.001)

Help

Process

RapidMiner Studio Core

Synopsis

The root operator which is the outer most operator of every process.

Description

19

Create split test design with <<Split Data>>

The screenshot displays the RapidMiner Studio interface with a workflow designed for a split test design. The workflow consists of the following operators:

- Read CSV**: The first operator in the process, connected to the input.
- Set Role**: Connected to the output of Read CSV, with 'exa' (example) and 'ori' (original) roles.
- Split Data**: Connected to the output of Set Role, with 'exa' and 'par' (partition) roles.
- Linear Regression**: The final operator, connected to the output of Split Data. It has 'tra' (training) and 'mod' (model) roles, and 'exa' and 'wei' (weights) roles.

Annotations and Callouts:

- 1**: A blue circle with the number 1, pointing to the 'Linear Regression' operator in the 'Repository' pane.
- 2**: A blue circle with the number 2, pointing to the 'Linear Regression' operator in the 'Process' pane.
- Use default parameter setting**: A text box pointing to the 'Linear Regression' operator in the 'Process' pane.
- Add <<Linear Regression>>**: A text box pointing to the 'Linear Regression' operator in the 'Repository' pane.

Parameters (Linear Regression):

- feature selection**: M5 prime
- eliminate colinear features**: ☒
- min tolerance**: 0.05
- use bias**: ☒
- ridge**: 1.0E-8
- [Hide advanced parameters](#)

Help (Linear Regression):

- Linear Regression**: RapidMiner Studio Core
- Tags**: Supervised, Classification, Regression, Model, Least squares, Ordinary, Ridge, Ols, Glm, Generalized, Functions
- Synopsis**: This operator calculates a linear

Exercise 1: Linear Regression with Rapidminer

Parameters of <<Linear Regression>>

- **feature selection**
 - Not all attributes are useful for linear regression.
 - Traditionally, forward selection (FS) or backward elimination (BE) are used to select features.
 - If you decide not to use these wrapper operators, you may use the ones which are bundled with multiple linear regression operator.
 - M5 prime, greedy, T-Test, iterative T-Test
- **eliminate collinear features**
 - This parameter indicates if the algorithm should try to delete collinear features during the regression or not.
 - min tolerance: This parameter is only available when the eliminate collinear features parameter is set to true. It specifies the minimum tolerance for eliminating collinear features.
- **use bias**
 - This parameter indicates if an intercept value should be calculated or not.
- **ridge**
 - This parameter specifies the ridge parameter for using in ridge regression.

Apply the model to the test dataset & Set regression performance measures

The screenshot displays the RapidMiner Studio interface with a workflow designed to apply a trained model to a test dataset and evaluate its performance using regression measures.

Workflow Steps:

- Read CSV:** Imports data from a file.
- Set Role:** Assigns roles to the data columns.
- Split Data:** Divides the data into training and testing sets.
- Linear Regression:** Trains a linear regression model on the training data.
- Apply Model:** Applies the trained model to the test dataset. A callout box labeled "Apply model to test dataset" points to this operator.
- Performance:** Calculates regression performance measures (e.g., lab, per, exa, wei). A callout box labeled "Measure regression performance" points to this operator.






Repository: Shows the data source structure, including Training Resources, Samples, Community Samples, Keras Samples, DB (Legacy), and Local Repository.

Operators: A search for "Perfro" (likely "Performance") shows several operators under the "Predictive" category, with "Performance (Regression)" selected.


Help: The "Apply Model" operator help is displayed, stating: "This Operator applies a model on an ExampleSet."

Footer: Business Data Mining ©Kim Yang Sok


Run the analysis process and check the results - Model





Views: Design **Results** Turbo Prep Auto Model Deployments


 All Studio ▼

Result History


 **LinearRegression (Linear Regression)** ×

 ExampleSet (Apply Model) ×


 PerformanceVector (Performance) ×



Data



Description



Annotations






Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
fixed acidity	0.018	0.020	0.038	0.956	0.921	0.357	
volatile acidity	-1.239	0.143	-0.269	0.826	-8.653	0	****
citric acid	-0.250	0.179	-0.060	0.815	-1.402	0.161	
residual sugar	0.021	0.015	0.037	1.000	1.471	0.142	
chlorides	-2.013	0.487	-0.122	0.994	-4.138	0.000	****
total sulfur dioxide	-0.002	0.001	-0.072	0.961	-2.714	0.007	***
pH	-0.520	0.186	-0.100	0.998	-2.800	0.005	***
sulphates	0.899	0.133	0.187	0.971	6.767	0.000	****
alcohol	0.301	0.021	0.396	0.883	14.474	0	****
(Intercept)	4.407	0.728	?	?	6.058	0.0	****

Check p-value and
remove attribute p-
value > 0.05


Business Data Mining ©Kim Yang Sok

23

Run the analysis process and check the results - Examples



Views: Design **Results** Turbo Prep Auto Model Deployments

 All Studio ▼

Result History

Regression (Linear Regres

ExampleSet (Apply Model)

PerformanceVector (Performance)

Data

Statistics

Visualizations

Annotations

Prep

Auto Mod

Real Quality

Predicted Quality

Filter (480 / 480 examples): all ▼

Row No.	quality	prediction(q...	fixed acidity	volatile acidity	citric acid	residual sug...	chlorides	free sulfur d...	total sulfur d...	density	pH
1	5	5.014	7.400	0.700	0	1.900	0.076	11	34	0.998	3.510
2	5	5.100	7.800	0.880	0	2.600	0.098	25	67	0.997	3.200
3	5	5.014	7.400	0.700	0	1.900	0.076	11	34	0.998	3.510
4	5	5.053	7.400	0.660	0	1.800	0.075	13	40	0.998	3.510
5	5	5.115	7.900	0.600	0.060	1.600	0.069	15	59	0.996	3.300
6	7	5.318	7.800	0.580	0.020	2	0.073	9	18	0.997	3.360
7	5	5.784	7.500	0.500	0.360	6.100	0.071	17	102	0.998	3.350
8	5	5.149	8.900	0.620	0.180	3.800	0.176	52	145	0.999	3.160
9	5	5.195	8.900	0.620	0.190	3.900	0.170	51	148	0.999	3.170
10	5	5.324	8.100	0.560	0.280	1.700	0.368	16	56	0.997	3.110
11	5	5.755	7.900	0.430	0.210	1.600	0.106	10	37	0.997	3.170
12	5	5.351	8.500	0.490	0.110	2.300	0.084	9	67	0.997	3.170
13	6	5.518	6.900	0.400	0.140	2.400	0.085	21	40	0.997	3.430
14	6	5.326	7.800	0.645	0	2	0.082	8	16	0.996	3.380

ExampleSet (480 examples, 2 special attributes, 11 regular attributes)

Business Data Mining ©Kim Yang Sok

24

Run the analysis process and check the results - Performance

//Local Repository/Business Data Mining/Lecture 8 Linear Regression/Modeling with Linear Regression* - RapidMiner Studio Educational 9.3.001 @ DESKTOP-TPCESUJ

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model

Find data, operators...etc All Studio

Result History LinearRegression (Linear Regression) PerformanceVector (Performance)

PerformanceVector

PerformanceVector:

- root_mean_squared_error: 0.640 +/- 0.000
- absolute_error: 0.496 +/- 0.404
- relative_error: 9.02% +/- 8.29%

Performance

Description

Annotations

Exercise 2:






Linear Regression after Removing Insignificant Attributes

Exercise 2: Linear Regression after Removing Insignificant Attributes

Task & Process


- **Task**
 - Build a model after removing insignificant attributes
- **Process**
 - Load “red wine” dataset
 - Add Select Attributes operator and remove insignificant attributes manually
 - Learn a model with the training datasets and apply it to the testing dataset
 - Run the analysis process and check the performance results


Run the analysis process and check the results - Model





Views: Design **Results** Turbo Prep Auto Model Deployments


Result History


LinearRegression (Linear Regression) 

ExampleSet (Apply Model) 

PerformanceVector (Performance) 


Data


Description


Annotations

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
fixed acidity	0.018	0.020	0.038	0.956	0.921	0.357	
volatile acidity	-1.239	0.143	-0.269	0.826	-8.653	0	****
citric acid	-0.250	0.179	-0.060	0.815	-1.402	0.161	
residual sugar	0.021	0.015	0.037	1.000	1.471	0.142	
chlorides	-2.013	0.487	-0.122	0.994	-4.138	0.000	****
total sulfur dioxide	-0.002	0.001	-0.072	0.961	-2.714	0.007	***
pH	-0.520	0.186	-0.100	0.998	-2.800	0.005	***
sulphates	0.899	0.133	0.187	0.971	6.767	0.000	****
alcohol	0.301	0.021	0.396	0.883	14.474	0	****
(Intercept)	4.407	0.728	?	?	6.058	0.000	****

Remove attributes
with p-value > 0.05

Business Data Mining ©Kim Yang Sok

28

Update process and run again!

Views: Design Results Turbo Prep Auto Model

Find data, operators...etc

All Studio

Repository

- Import Data
- Business Data Mining (admin)
 - Lecture 5 Data Preparation (admin)
 - Lecture 6 Modeling (admin)
 - data (admin)
 - process (admin)
 - Lecture 7 Test Design (admin)
 - Lecture 8 Linear Regression (admin)
 - Data Loading Practice 2 (admin - v)
 - Handling Missing Data (admin - v)

Process

Process

100%

Read CSV

Set Role

Split Data

Linear Regression

Apply Model

Performance

Select Attributes

Select attributes except 'fixed acidity', 'citric acid', and 'residual sugar'

Parameters

Apply Model

application paramet... Edit List (0)...

☐ create view

[Hide advanced parameters](#)

[Change compatibility \(9.3.001\)](#)

Operators

Select Attribute

- Selection (4)
 - Select Attributes
 - Remove Attribute Range
 - Remove Useless Attributes
 - Remove Correlated Attributes
- Modeling (2)
 - Optimization (2)
 - Automatic Feature Engineering

No results were found.

Help

Apply Model

RapidMiner Studio Core

Tags: [Predict](#), [Predictions](#), [Forecasts](#), [Scores](#), [Scoring](#), [Trained](#), [Test](#)

Synopsis

This Operator applies a model on an ExampleSet.

[Jump to Tutorial Process](#)

Business Data Mining ©Kim Yang Sok

29

Run the analysis process and check the results - Model

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
fixed acidity	0.018	0.020	0.038	0.956	0.921	0.357	
volatile acidity	-1.239	0.143	-0.269	0.826	-8.653	0	
citric acid	-0.250	0.179	-0.060	0.815	-1.402	0.161	
residual sugar	0.021	0.015	0.037	1.000	1.471	0.142	
chlorides	-2.013	0.487	-0.122	0.994	-4.138	0.000	****
total sulfur dioxide	-0.002	0.001	-0.072	0.961	-2.714	0.007	***
pH	-0.520	0.186	-0.100	0.998	-2.800	0.005	***
sulphates	0.899	0.133	0.187	0.971	6.767	0.000	****
alcohol	0.301	0.021	0.396	0.883	14.474	0	
(Intercept)	4.407	0.728	?	?	6.058	0.000	

PerformanceVector:
root_mean_squared_error: 0.640 +/- 0.000
absolute_error: 0.496 +/- 0.404
relative_error: 9.02% +/- 8.29%

PerformanceVector:
root_mean_squared_error: 0.638 +/- 0.000
absolute_error: 0.496 +/- 0.402
relative_error: 9.00% +/- 8.25%



Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
volatile acidity	-1.127	0.121	-0.244	0.880	-9.289	0	****
chlorides	-2.159	0.462	-0.130	0.997	-4.673	0.000	****
total sulfur dioxide	-0.002	0.001	-0.073	0.968	-2.977	0.003	***
pH	-0.548	0.140	-0.105	1.000	-3.924	0.000	****
sulphates	0.885	0.132	0.184	0.971	6.715	0.000	****
alcohol	0.298	0.020	0.392	0.902	14.790	0	****
(Intercept)	4.635	0.476	?	?	9.745	0	****



Data



Description



Annotations



Views:

Design

Results

Turbo Prep

Auto Model

Find data, operators...etc



All Studio

Result History LinearRegression (Linear Regression) ExampleSet (Apply Model) PerformanceVector (Performance)

Open in Turbo Prep Auto Model

Filter (480 / 480 examples): all

Row No.	quality	prediction(q...	volatile acidity	chlorides	free sulfur d...	total sulfur d...	density	pH	sulphates	alcohol
1	5	4.996	0.700	0.076	11	34	0.998	3.510	0.560	9.400
2	5	5.081	0.880	0.098	25	67	0.997	3.200	0.680	9.800
3	5	4.996	0.700	0.076	11	34	0.998	3.510	0.560	9.400
4	5	5.032	0.660	0.075	13	40	0.998	3.300	0.560	9.400
5	5	5.105	0.600	0.069	15	59	0.996	3.300	0.460	9.400



Regression function

Quality = - 1.127 * volatile acidity
- 2.159 * chlorides
- 0.002 * total sulfur dioxide
- 0.548 * pH
+ 0.885 * sulphates
+ 0.298 * alcohol
+ 4.635

	coefficient	attribute value	multiplication
volatile acidity	-1.127	0.7	-0.789
chlorides	-2.159	0.076	-0.164
total sulfur dioxide	-0.002	34	-0.068
pH	-0.548	3.51	-1.923
sulphates	0.885	0.56	3.106
alcohol	0.298	9.4	0.167
Intercept			4.635
		prediction	4.964

16	5	5.340	0.320	0.103	13	50	0.996	3.380	0.550	9.200
----	---	-------	-------	-------	----	----	-------	-------	-------	-------

ExampleSet (480 examples, 2 special attributes, 8 regular attributes)

Exercise 3:

Linear Regression with Variance Filter

- **Variance thresholds remove features whose values don't change much from observation to observation (i.e. their variance falls below a threshold). These features provide little value.**
- **For example, if you had a public health dataset where 96% of observations were for 35-year-old men, then the 'Age' and 'Gender' features can be eliminated without a major loss in information.**
- **Because variance is dependent on scale, you should always normalize your features first.**

- **Strengths**

- Applying variance thresholds is based on solid intuition: features that don't change much also don't add much information.
- This is an easy and relatively safe way to reduce dimensionality at the start of your modeling process.

- **Weaknesses**

- If your problem does require dimensionality reduction, applying variance thresholds is rarely sufficient.
- Furthermore, you must manually set or tune a variance threshold, which could be tricky.
- It is better to start with a conservative (i.e. lower) threshold.



Views:

Design

Results

Turbo Prep

Auto Model

Deployments

Find data, open

Change threshold

Repository

+ Import Data

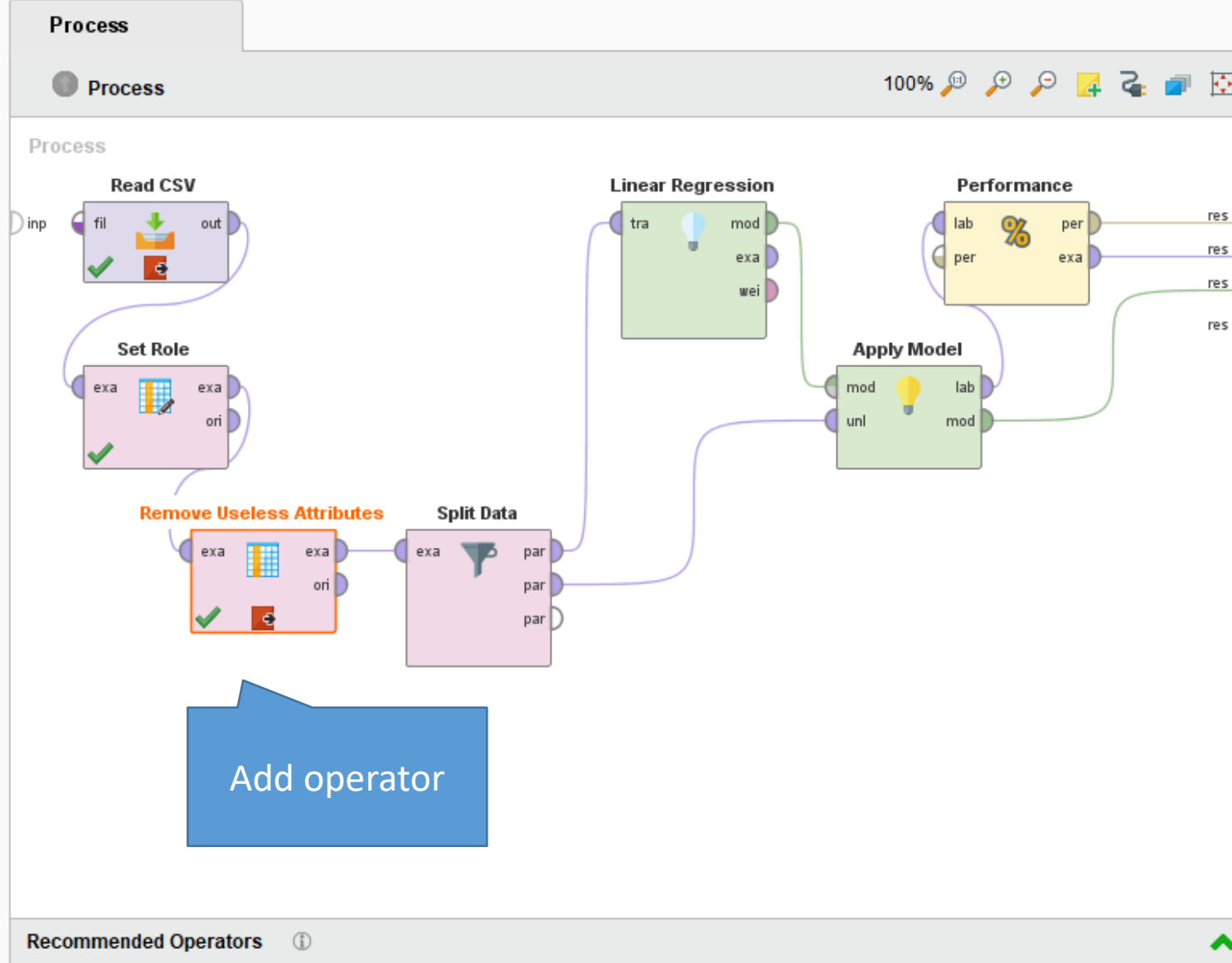
- Training Resources (connected)
- Samples
- Community Samples (connected)
- Keras Samples
- DB (Legacy)
- Local Repository (admin)

Operators

Remove

- Blending (6)
 - Attributes (4)
 - Selection (4)
 - Select Attributes
 - Remove Attribute Range
 - Remove Useless Attributes
 - Remove Correlated Attributes

No results were found.



Parameters

Remove Useless Attributes

numerical min devi... 0.2

nominal useless a... 1.0

☐ nominal remove id like

nominal useless b... 0.0

Help

Remove Useless Attributes

RapidMiner Studio Core

Tags: [Filter](#), [Keep](#), [Remove](#), [Drop](#), [Delete](#), [Columns](#), [Variables](#), [Features](#), [Feature Set](#), [Constant](#), [Deviation](#), [Variance](#), [Selection](#)

Exercise 4:

Linear Regression

with Correlation Filter

- Correlation thresholds remove features that are highly correlated with others (i.e. its values change very similarly to another's). These features provide redundant information.
- For example, if you had a real-estate dataset with 'Floor Area (Sq. Ft.)' and 'Floor Area (Sq. Meters)' as separate features, you can safely remove one of them.
- Which one should you remove? Well, you'd first calculate all pair-wise correlations. Then, if the correlation between a pair of features is above a given threshold, you'd remove the one that has larger mean absolute correlation with other features.

- **Strengths**

- Applying correlation thresholds is also based on solid intuition: similar features provide redundant information.
- Some algorithms are not robust to correlated features, so removing them can boost performance.

- **Weaknesses:**

- You must manually set or tune a correlation threshold, which can be tricky to do.
- In addition, if you set your threshold too low, you risk dropping useful information.
- Whenever possible, we prefer algorithms with built-in feature selection over correlation thresholds.
- Even for algorithms without built-in feature selection, Principal Component Analysis (PCA) is often a better alternative.



Views:

Design

Results

Turbo Prep

Auto Model

Deployments

Find data, open

Change threshold

Repository

+ Import Data

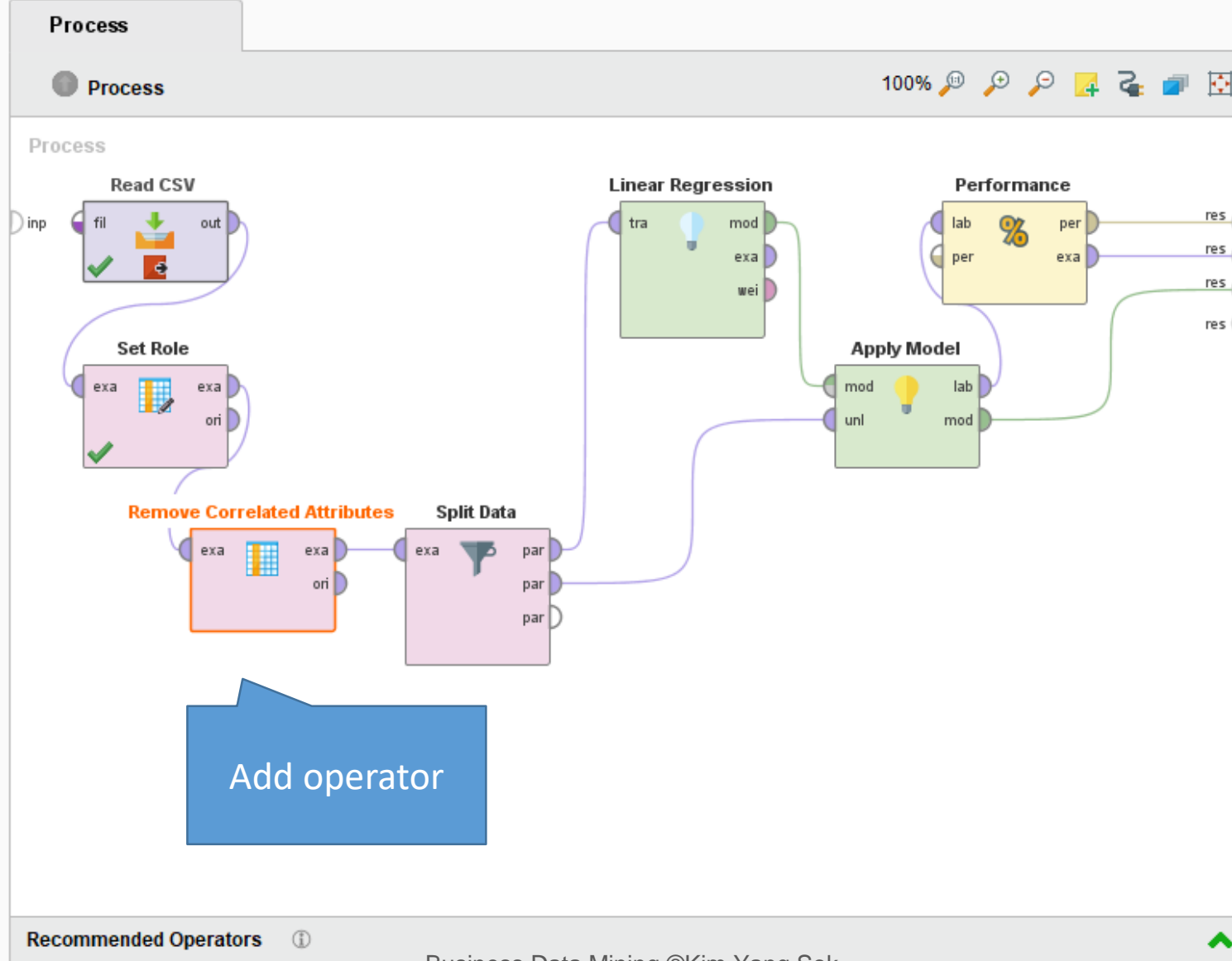
- Training Resources (connected)
- Samples
- Community Samples (connected)
- Keras Samples
- DB (Legacy)
- Local Repository (admin)

Operators

Remove

- Blending (6)
- Attributes (4)
- Selection (4)
- Select Attributes
- Remove Attribute Range
- Remove Useless Attributes
- Remove Correlated Attributes

No results were found.



Parameters

Remove Correlated Attributes

correlation 0.7

filter relation greater

attribute order original

☒ use absolute correlation

☐ use local random seed

[Hide advanced parameters](#)

☒ [Change compatibility \(9.5.000\)](#)

Help

Remove Correlated Attributes

RapidMiner Studio Core

Tags: [Filter](#), [Keep](#), [Remove](#), [Drop](#), [Delete](#), [Correlations](#), [Selection](#)

Synopsis

This operator removes

Exercise 5:

Linear Regression with Weight Filter

- **Filter feature selection methods apply a statistical measure to assign a scoring to each feature.**
- **The features are ranked by the score and either selected to be kept or removed from the dataset.**
- **The methods are often univariate and consider the feature independently, or with regard to the dependent variable.**
- **Some examples of some filter methods include the Chi squared test, information gain and correlation coefficient scores.**



Views:

Design

Results

Turbo Prep

Auto Model

Deployments

Find data, operators...etc

All Studio

Repository

+ Import Data

Training Resources (connected)

Samples

Community Samples (connected)

Keras Samples

DB (Legacy)

Local Repository (admin)

Operators

Select by

Blending (2)

Attributes (2)

Selection (2)

Select by Weights

Select by Random

No results were found.

Process

Process

100%

Process

Read CSV



Set Role



Weight by Correlation



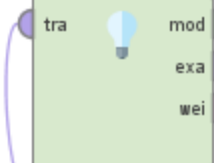
Select by Weights



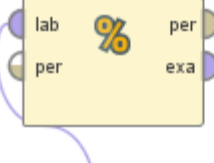
Split Data



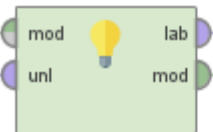
Linear Regression



Performance



Apply Model



Filter by weights

Calculate correlation between attributes and label

Parameters

Select by Weights

weight relation greater equals

weight 0.05

deselect unknown

use

Set threshold

Hide advanced parameters

Help

Select by Weights

RapidMiner Studio Core

Tags: Weighting, Importance, Influence, Significance, Factors, Relevance, Thresholds, Selection

Synopsis

This operator select attributes of an issue

Exercise 6:

Linear Regression with Forward Selection using Rapidminer

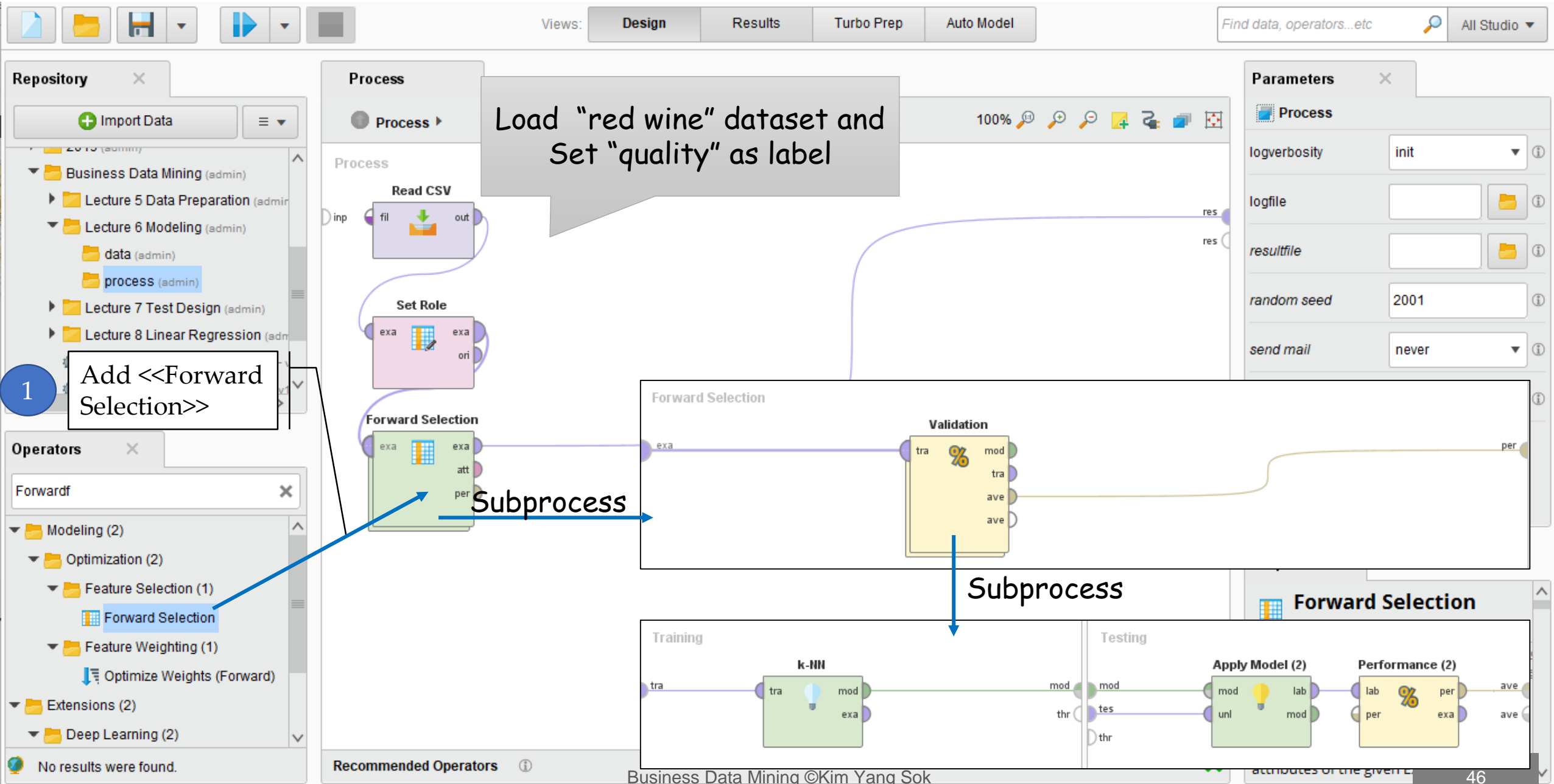
- Stepwise search is a supervised feature selection method based on sequential search, and it has two flavors: forward and backward.
- Forward stepwise search
 - Start without any features.
 - Then, you'd train a 1-feature model using each of your candidate features and keep the version with the best performance.
 - You'd continue adding features, one at a time, until your performance improvements stall.
- Backward stepwise search
 - Start with all features in your model and then remove one at a time until performance starts to drop substantially.
- Despite many textbooks listing stepwise search as a valid option, it almost always **underperforms other supervised methods** such as regularization.

Exercise 6: Linear Regression with Feature Selection using Rapidminer

Task & Process

- **Task**
 - After loading data, build a linear regression model with feature selection.
- **Process**
 - Load “red wine” dataset and Set “quality” as label
 - Select attributes with forward selection approach
 - Perform a linear regression modeling with the split test design
 - Run the analysis process and evaluate analysis results

Select attributes with forward selection approach



Perform a linear regression modeling with the split test design

Repository

- Import Data
- 2013 Learning
 - Business Data Mining (admin)
 - Lecture 5 Data Preparation (admin)
 - Lecture 6 Modeling (admin)
 - data (admin)
 - process (admin)
 - Lecture 7 Test Design (admin)
 - Lecture 8 Linear Regression (admin)
 - Data Loading Practice 2 (admin)
 - Handling Missing Data (admin)

Process

Process

100%

Read CSV

Set Role

Forward Selection

Split Data

Linear Regression

Apply Model

Performance

Parameters

Forward Selection

maximal number of ... 10

speculative rounds 0

stopping behavior without increase

Operators

Performance

- Validation (19)
 - Performance (17)
 - Predictive (7)
 - Performance (Classification)
 - Performance (Binominal Clas
 - Performance (Regression)
 - Performance (Costs)

Help

Forward Selection

RapidMiner Studio Core

Tags: Iterate, Iteration, Weighting, Importance, Influence, Significance, Factors, Relevance, Feature Selection

Synopsis






This operator selects the most relevant attributes of the given E

Callout: Set 'Breakpoint After' for <<Set Role>> and <<Forward Selection>> to check selected attributes

Recommended Operators

Business Data Mining ©Kim Yang Sok

Run the analysis process and evaluate analysis results


Views:

Design

Results

Turbo Prep

Auto Model



All Studio ▾

Result History


ExampleSet (Set Role)


Open in  Turbo Prep  Auto Model Filter (1,599 / 1,599 examples): all

Row No.	quality	fixed acidity	volatile acidity	citric acid	residual sug...	chlorides	free sulfur d...	total sulfur d...	density	pH	sulphates	alcohol
1	5	7.400	0.700	0	1.900	0.076	11	34	0.998	3.510	0.560	9.400
2	5	7.800	0.880	0	2.600	0.098	25	67	0.997	3.200	0.680	9.800
3	5	7.800	0.760	0.040	2.300	0.092	15	54	0.997	3.260	0.650	9.800

\\Local Repository\Business Data Mining\Lecture 8 Linear Regression\Modeling with Linear Regression Using Forward Selection* – RapidMiner Studio Educational 9.3.001 @ DESKTOP-TPCESUJ




7
8 Result History ExampleSet (Forward Selection) X




Annotations 9 10 Open in Turbo Prep Auto Model Filter (1,599 /

Row No.	quality	sulphates	alcohol	pH	volatile acidity
1	5	0.560	9.400	3.510	0.700
2	5	0.680	9.800	3.200	0.880
3	5	0.650	9.800	3.260	0.760
4	6	0.580	9.800	3.160	0.280
5	5	0.560	9.400	3.510	0.700
6	5	0.560	9.400	3.510	0.660

After applying forward selection, 4 attributes are selected.

Run the analysis process and evaluate analysis results






Views:

Design

Results

Turbo Prep


Auto Model


Find data, operators...etc  All Studio ▼


Result History

LinearRegression (Linear Regression) ×

PerformanceVector (Performance) ×


Data


Description


Annotations

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
sulphates	0.560	0.115	0.124	0.945	4.868	0.000	****
alcohol	0.342	0.020	0.439	0.959	17.156	0	****
pH	-0.392	0.136	-0.076	1.000	-2.880	0.004	***
volatile acidity	-1.251	0.115	-0.282	0.910	-10.843	0	****
(Intercept)	3.665	0.449	?	?	8.167	0.000	****

PerformanceVector:

root_mean_squared_error: 0.693 +/- 0.000

absolute_error: 0.540 +/- 0.434

relative_error: 9.93% +/- 9.50%

Business Data Mining ©Kim Yang Sok

49

Exercise 7:

Linear Regression with Backward Elimination using Rapidminer

Exercise 7: Linear Regression with Feature Selection using Rapidminer

Task & Process

- **Task**
 - After loading data, build a linear regression model with **backward elimination**.
- **Process**
 - Load “red wine” dataset and Set “quality” as label
 - Select attributes with **backward elimination** approach
 - Perform a linear regression modeling with the split test design
 - Run the analysis process and evaluate analysis results

Perform a linear regression modeling with the split test design

The screenshot displays the RapidMiner Studio interface with a process design canvas. The process flow is as follows:

- Read CSV**: Loads the 'red wine' dataset.
- Set Role**: Sets the 'quality' attribute as the label.
- Backward Elimination**: Performs feature selection using backward elimination.
- Split Data**: Splits the data into training and testing sets.
- Linear Regression**: Trains a linear regression model on the training data.
- Apply Model**: Applies the trained model to the test data.
- Performance**: Evaluates the model's performance.

Annotations and callouts provide additional context:

- Load 'red wine' dataset**: Points to the 'Read CSV' operator.
- Set 'quality' as label**: Points to the 'Set Role' operator.
- Use <<Backward Elimination>> with k-NN. Analysis design of this operator is the same as that of <<Forward Selection>>**: Points to the 'Backward Elimination' operator.
- Linear regression process design**: Points to the overall process flow.






The **Parameters** panel on the right shows the configuration for the 'Backward Elimination' operator:

- maximal number of ...: 10
- speculative rounds: 0
- stopping behavior: with decrease


The **Operators** panel on the left shows the search results for the 'Backward Elimination' operator.


Business Data Mining ©Kim Yang Sok


Run the analysis process and evaluate analysis results





Views: Design Results Turbo Prep Auto Model


Find data, operators...etc 

All Studio 


Result History ExampleSet (Set Role) 


Data






Open in  Turbo Prep  Auto Model

Filter (1,599 / 1,599 examples): all 


Row No.	quality	fixed acidity	volatile acidity	citric acid	residual sug...	chlorides	free sulfur d...	total sulfur d...	density	pH	sulphates	alcohol
1	5	7.400	0.700	0	1.900	0.076	11	34	0.998	3.510	0.560	9.400
2	5	7.800	0.880	0	2.600	0.098	25	67	0.997	3.200	0.680	9.800


 //Local Repository/Business Data Mining/Lecture 8 Linear Regression/Modeling with Linear Regression Using Backward Elimination – RapidMiner Studio Educational 9.3.001 @ DESKTOP-TPCESUJ


File Edit Process View Connections Settings Extensions Help






Views: Design Results Turbo Prep Auto Model


Find data, operators...etc 

All Studio 

Result History ExampleSet (Backward Elimination) 


Data

Open in  Turbo Prep  Auto Model

Filter (1,599 / 1,599 examples): all 






Row No.	quality	fixed acidity	volatile acidity	citric acid	residual sug...	chlorides	total sulfur d...	density	pH	sulphates	alcohol
1	5	7.400	0.700	0	1.900	0.076	34	0.998	3.510	0.560	9.400
2	5	7.800	0.880	0	2.600	0.098	67	0.997	3.200	0.680	9.800
3	5	7.800	0.760	0.040	2.300	0.092	54	0.996	3.300	0.460	9.800
4	6	11.200	0.280	0.560	1.900	0.075	60	0.996	3.300	0.460	9.800
5	5	7.400	0.700	0	1.900	0.076	34	0.996	3.300	0.460	9.400
6	5	7.400	0.660	0	1.800	0.075	40	0.996	3.300	0.460	9.400
7	5	7.900	0.600	0.060	1.600	0.069	59	0.996	3.300	0.460	9.400
8	7	7.300	0.650	0	1.200	0.065	21	0.995	3.390	0.470	10

After applying backward elimination, 10 attributes are selected except 'free sulfur dioxide'.


Business Data Mining ©Kim Yang Sok

53


Run the analysis process and evaluate analysis results





Views: Design **Results** Turbo Prep Auto Model


 All Studio ▾


Result History

 LinearRegression (Linear Regression) ×

 PerformanceVector (Performance) ×


Data


Description


Annotations

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
fixed acidity	-0.008	0.017	-0.016	0.950	-0.468	0.640	
volatile acidity	-1.202	0.120	-0.263	0.859	-9.999	0	****
residual sugar	0.016	0.014	0.027	1.000	1.086	0.278	
chlorides	-1.786	0.458	-0.109	0.994	-3.904	0.000	****
total sulfur dioxide	-0.002	0.001	-0.089	0.973	-3.516	0.000	****
pH	-0.482	0.193	-0.089	0.998	-2.491	0.013	**
sulphates	0.879	0.140	0.174	0.949	6.294	0.000	****
alcohol	0.291	0.020	0.386	0.871	14.598	0	****
(Intercept)	4.517	0.755	?	?	5.986	0.000	****

PerformanceVector:

root_mean_squared_error: 0.659 +/- 0.000

absolute_error: 0.509 +/- 0.419

relative_error: 8.99% +/- 7.92%

Exercise 8:

Linear Regression with

Model Explainer

Exercise 8: Linear Regression with Model Explainer

Task & Process

- **Task**
 - Which attributes play the largest role when making a prediction
- **Process**
 - Load “red wine” dataset and Set “quality” as label
 - Split data into training and testing datasets
 - Learn a linear regression model with training dataset
 - Add Explain Predictions operator and provide the model learned and testing dataset
 - Run the analysis process and examine prediction results

-



Views:

Design

Results

Turbo Prep

Auto Model

Deployments

Find data, operators...etc

All Studio

Repository

Import Data

Training Resources (connected)

Samples

Community Samples (connected)

Keras Samples

DB (Legacy)

Local Repository (admin)

Operators

Split

Sampling (1)

Split Data

Values (1)

Split

Validation (3)

We found "Information Extraction" in the Marketplace. [Show me!](#)

Process

Process

100%

Process

Read CSV

Linear Regression

Multiply

Set Role

Split Data

Explain Predictions

Parameters

Explain Predictions

maximal explaining... 3

local sample size 500

☐ only create predictions

☐ normalize global weights

☒ sort weights

sort direction descending

[Hide advanced parameters](#)

Help

Explain Predictions

Model Simulator

Tags: [Explanations](#), [Supporting](#), [Contradicting](#), [LIME](#), [Narrative](#), [Scoring](#)

Synopsis

This operator identifies the attributes that play the largest role when

Recommended Operators

Views Turbo Prep Auto Model Deployments Find data, operators...etc All Studio

ExampleSet (Explain Predictions) Negative Impact Positive Impact regression)

Row No.	quality	prediction(q...	fixed acidity	volatile acidity	citric...	resi...	chlor...	fre...	tot...	den...	pH	sulp...	alcohol
1	5	5.014	7.400	0.700	0	1.900	0.076	11	34	0.998	3.510	0.560	9.400
2	5	5.100	7.800	0.880	0	2.600	0.098	25	67	0.997	3.200	0.680	9.800
3	5	5.014	7.400	0.700	0	1.900	0.076	11	34	0.998	3.510	0.560	9.400
4	5	5.053	7.400	0.660	0	1.800	0.075	13	40	0.998	3.510	0.560	9.400
5	5	5.115	7.900	0.600	0.060	1.600	0.069	15	59	0.996	3.300	0.460	9.400
6	7	5.318	7.800	0.580	0.020	2	0.073	9	18	0.997	3.360	0.570	9.500
7	5	5.784	7.500	0.500	0.360	6.100	0.071	17	102	0.998	3.350	0.800	10.500
8	5	5.149	8.900	0.620	0.180	3.800	0.176	52	145	0.999	3.160	0.880	9.200
9	5	5.195	8.900	0.620	0.190	3.900	0.170	51	148	0.999	3.170	0.930	9.200
10	5	5.324	8.100	0.560	0.280	1.700	0.368	16	56	0.997	3.110	1.280	9.300
11	5	5.755	7.900	0.430	0.210	1.600	0.106	10	37	0.997	3.170	0.910	9.500
12	5	5.351	8.500	0.490	0.110	2.300	0.084	9	67	0.997	3.170	0.530	9.400
13	6	5.518	6.900	0.400	0.140	2.400	0.085	21	40	0.997	3.430	0.630	9.700
14	6	5.326	7.800	0.645	0	2	0.082	8	16	0.996	3.380	0.590	9.800
15	5	5.298	8.300	0.655	0.120	2.300	0.083	15	113	0.997	3.170	0.660	9.800
16	5	5.295	5.200	0.320	0.250	1.800	0.103	13	50	0.996	3.380	0.550	9.200

Data Statistics Visualizations Annotations	Open in Turbo Prep Auto Model		Filter (480 / 480 examples): all			
	Row No.	quality	prediction(q...	Support Prediction	Contradict Prediction	fixed acidity
	1	5	5.014	alcohol = 9.400 (0.586); sulphates = 0.560 (0.498); residual sugar = ...	volatile acidity = 0.700 (-0.583); chlorides = 0.076 (-0.343); pH = 3....	7.400
	2	5	5.100	alcohol = 9.800 (0.586); sulphates = 0.680 (0.498); residual sugar = ...	volatile acidity = 0.880 (-0.583); chlorides = 0.098 (-0.343); pH = 3....	7.800
	3	5	5.014	alcohol = 9.400 (0.586); sulphates = 0.560 (0.498); residual sugar = ...	volatile acidity = 0.700 (-0.583); chlorides = 0.076 (-0.343); pH = 3....	7.400
	4	5	5.053	alcohol = 9.400 (0.586); sulphates = 0.560 (0.498); residual sugar = ...	volatile acidity = 0.660 (-0.583); chlorides = 0.075 (-0.343); pH = 3....	7.400
	5	5	5.115	alcohol = 9.400 (0.586); sulphates = 0.460 (0.498); residual sugar = ...	volatile acidity = 0.600 (-0.583); chlorides = 0.069 (-0.343); pH = 3....	7.900
	6	7	5.318	alcohol = 9.500 (0.586); sulphates = 0.570 (0.498); residual sugar = ...	volatile acidity = 0.580 (-0.583); chlorides = 0.073 (-0.343); pH = 3....	7.800
	7	5	5.784	alcohol = 10.500 (0.586); sulphates = 0.800 (0.498); residual sugar = ...	volatile acidity = 0.500 (-0.583); chlorides = 0.071 (-0.343); pH = 3....	7.500
	8	5	5.149	alcohol = 9.200 (0.586); sulphates = 0.880 (0.498); residual sugar = ...	volatile acidity = 0.620 (-0.583); chlorides = 0.176 (-0.343); pH = 3....	8.900
	9	5	5.195	alcohol = 9.200 (0.586); sulphates = 0.930 (0.498); residual sugar = ...	volatile acidity = 0.620 (-0.583); chlorides = 0.170 (-0.343); pH = 3....	8.900
	10	5	5.324	alcohol = 9.300 (0.586); sulphates = 1.280 (0.498); residual sugar = ...	volatile acidity = 0.560 (-0.583); chlorides = 0.368 (-0.343); pH = 3....	8.100
	11	5	5.755	alcohol = 9.500 (0.586); sulphates = 0.910 (0.498); residual sugar = ...	volatile acidity = 0.430 (-0.583); chlorides = 0.106 (-0.343); pH = 3....	7.900
	12	5	5.351	alcohol = 9.400 (0.586); sulphates = 0.530 (0.498); residual sugar = ...	volatile acidity = 0.490 (-0.583); chlorides = 0.084 (-0.343); pH = 3....	8.500
	13	6	5.518	alcohol = 9.700 (0.586); sulphates = 0.630 (0.498); residual sugar = ...	volatile acidity = 0.400 (-0.583); chlorides = 0.085 (-0.343); pH = 3....	6.900

ExampleSet (480 examples, 4 special attributes, 11 regular attributes)

Open in Turbo Prep Auto Model

Filter (5,280 / 5,280 examples): all

Row No.	Row No	Name	Value	Importance
1	1	fixed acidity	7.4	0.032
2	1	volatile acidity	0.7	-0.583
3	1	citric acid	0.0	-0.042
4	1	residual sugar	1.9	0.085
5	1	chlorides	0.076	-0.343
6	1	free sulfur dio...	11.0	0.033
7	1	total sulfur di...	34.0	-0.195
8	1	density	0.9978	0.019
9	1	pH	3.51	-0.238
10	1	sulphates	0.56	0.498
11	1	alcohol	9.4	0.586
12	2	fixed acidity	7.8	0.032
13	2	volatile acidity	0.88	-0.583
14	2	citric acid	0.0	-0.042

ExampleSet (5,280 examples, 0 special attributes, 4 regular attributes)



Views:

Design

Results

Turbo Prep

Auto Model

Deployments

Find data, operators...etc



All Studio

ExampleSet (Explain Predictions)

! ExplainPredictionsIOObject (Explain Predictions)

LinearRegression (Linear Regression)

Result History

AttributeWeights (Explain Predictions)

ExampleSet (Explain Predictions)



Data



Weight
Visualizations



Annotations

attribute	weight
alcohol	0.586
sulphates	0.498
residual sugar	0.085
free sulfur dioxide	0.033
fixed acidity	0.032
density	0.019
volatile acidity	0
citric acid	0
chlorides	0
total sulfur dioxide	0
pH	0

Exercise 9:

Linear Regression with

Model Simulator

Exercise 9: Linear Regression with Model Simulator

Task & Process

- **Task**
 - Simulate the target value with the model
- **Process**
 - Load “red wine” dataset and Set “quality” as label
 - Split data into training and testing datasets
 - Learn a linear regression model with training dataset
 - Add Model Simulator operator and provide the model learned and testing dataset
 - Run the analysis process and examine simulation results

-



Views:

Design

Results

Turbo Prep

Auto Model

Deployments

Find data, operators...etc



All Studio

Repository

+ Import Data

- Training Resources (connected)
- Samples
- Community Samples (connected)
- Keras Samples
- DB (Legacy)
- Local Repository (admin)

Operators

Split

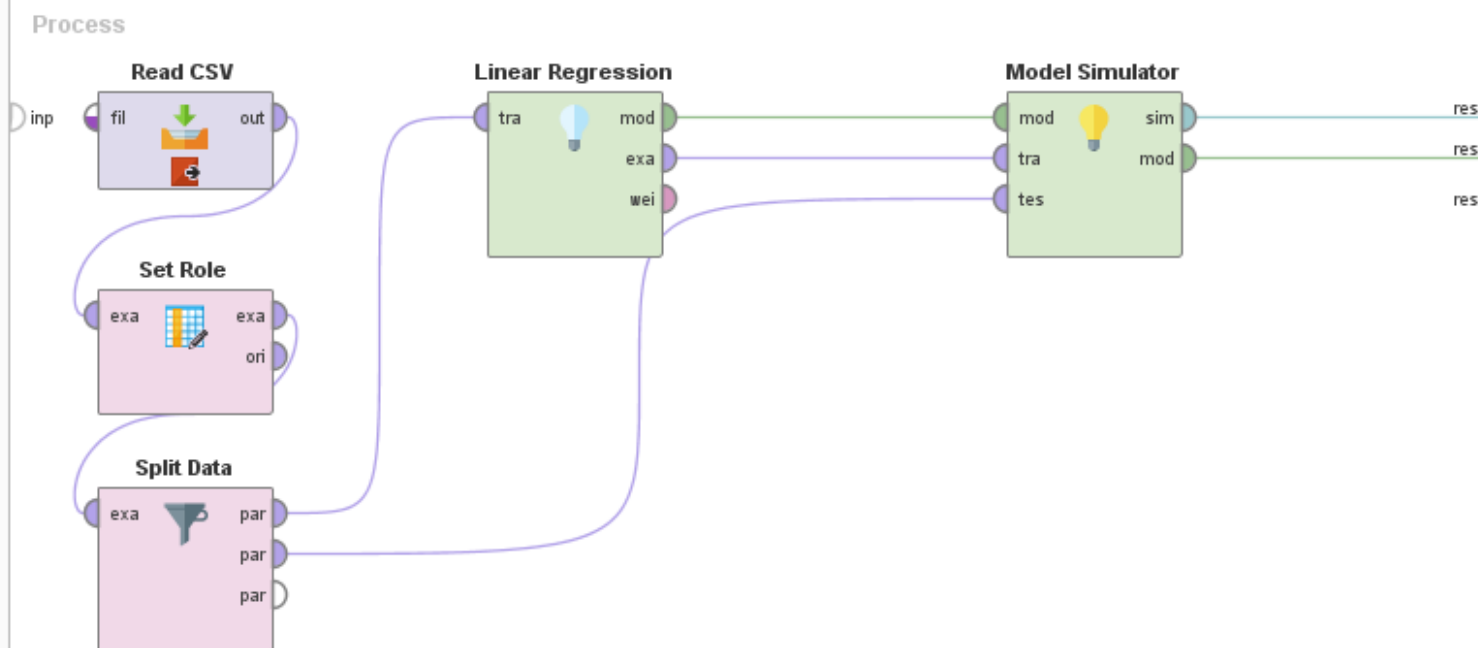
- Sampling (1)
 - Split Data
- Values (1)
 - Split
- Validation (3)

We found "Information Extraction" in the Marketplace. [Show me!](#)

Process

Process

100%



Recommended Operators

Parameters

Process

logverbosity init

logfile

resultfile

random seed 2001

send mail never

encoding SYSTEM

[Hide advanced parameters](#)

[Change compatibility \(9.5.000\)](#)

Help

Process

RapidMiner Studio Core

Synopsis

The root operator which is the outer most operator of every process.

Description



Input for Model

chlorides:	<input type="text" value="0.085"/>	①
alcohol:	<input type="text" value="9"/>	①
density:	<input type="text" value="0.997"/>	①
free sulfur dioxide:	<input type="text" value="16.082"/>	①
fixed acidity:	<input type="text" value="8.410"/>	①
total sulfur dioxide:	<input type="text" value="46.933"/>	①
volatile acidity:	<input type="text" value="0.534"/>	①
sulphates:	<input type="text" value="0.660"/>	①
citric acid:	<input type="text" value="0.264"/>	①
pH:	<input type="text" value="3.310"/>	①
residual sugar:	<input type="text" value="2.537"/>	①

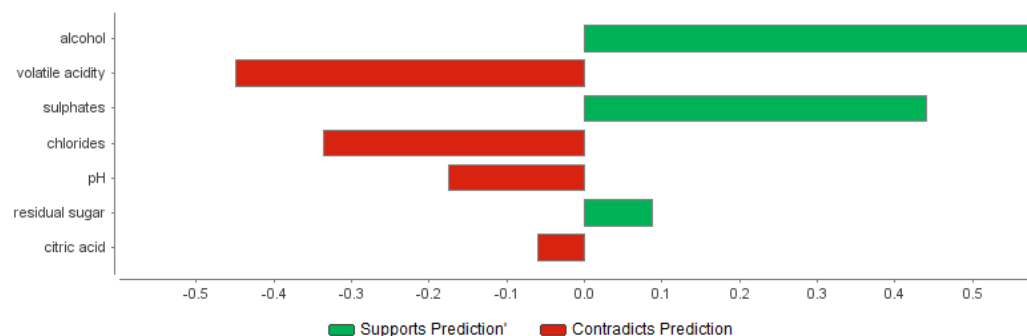
Optimize What is [this?](#)

Prediction

Prediction

5.218

Important Factors for Prediction



Interpretation

You can change values and see the quality!

See the model's reaction on the right. The prediction of the model is 5.218. The biggest support for this decision is coming from **alcohol**.

Distribution of Predictions

You need to provide a test data set to see a distribution of predictions.

Accuracy

Accuracy can not be calculated: no test data was provided.



ModelSimulatorI0Object (Model Simulator)

Result History

Model
Simulator

Input for Model

chlorides:



alcohol:



density:



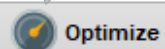
free sulfur dioxide:



fixed acidity:



Click here to find
optimized values



Optimize

What is [this](#)?

Find Optimal Input Settings

Define Targets Define Constraints Optimization Parameters Running Optimization

Let this optimization find input values to the prediction you desire. Define below if you want to maximize or minimize the forecast of the model. Or specify a specific value you would like to reach.

Direction for predictions: Maximize

Specific prediction to reach: 42

< Back Next >

sulphates

chlorides



ModelSimulatorI0Object (Model Simulator)

Result History

Model
Simulator

Input for Model

chlorides:

alcohol:

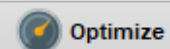
density:

free sulfur dioxide:

fixed acidity:

total sulfur dioxide:

volatile acidity:



What is [this?](#)

Find Optimal Input Settings

Define Targets Define Constraints Optimization Parameters Running Optimization

Often you need to define constraints on the model inputs so that only reasonable solutions are created. You can set global constraints which apply to all inputs or constant attributes.

Global Constraints

- ☒ Stay within 2 standard deviations from average (recommended)
- ☒ Stay above all attribute minimum
- ☒ Stay below all attribute maximum
- ☐ Stay above
- ☐ Stay below

Constant Attributes

+ Add new constant attribute

Back Next

Set constraints

Define Constant Attribute

Select an attribute on the left and set the desired constant value for this attribute on the right.

alcohol 9

Cancel Ok



ModelSimulatorIOObject (Model Simulator)

Result History

Model
Simulator

Input for Model

chlorides:



alcohol:



density:



free sulfur dioxide:



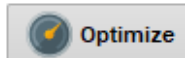
fixed acidity:



total sulfur dioxide:



volatile acidity:



Optimize

What is [this](#)?

Find Optimal Input Settings

Define Targets

Define Constraints

Optimization Parameters

Running Optimization

Finally, you can define how long the optimization will run. No limit means that the optimization runs until it finds the optimal inputs. Or specify a maximum time or the maximum number of generations at which point the best result so far will be delivered.

☒ No Limit

☐ Time Limit

Seconds:

10

☐ Generation Limit

Generations:

500

Population:

50



Back



Run

sulphates

chlorides



ModelSimulator10Object (Model Simulator)

Result History

Model
Simulator

Input for Model

chlorides:



alcohol:



density:



free sulfur dioxide:



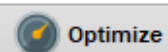
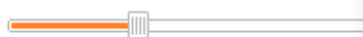
fixed acidity:



total sulfur dioxide:



volatile acidity:



Optimize

What is [this](#)?

Find Optimal Input Settings

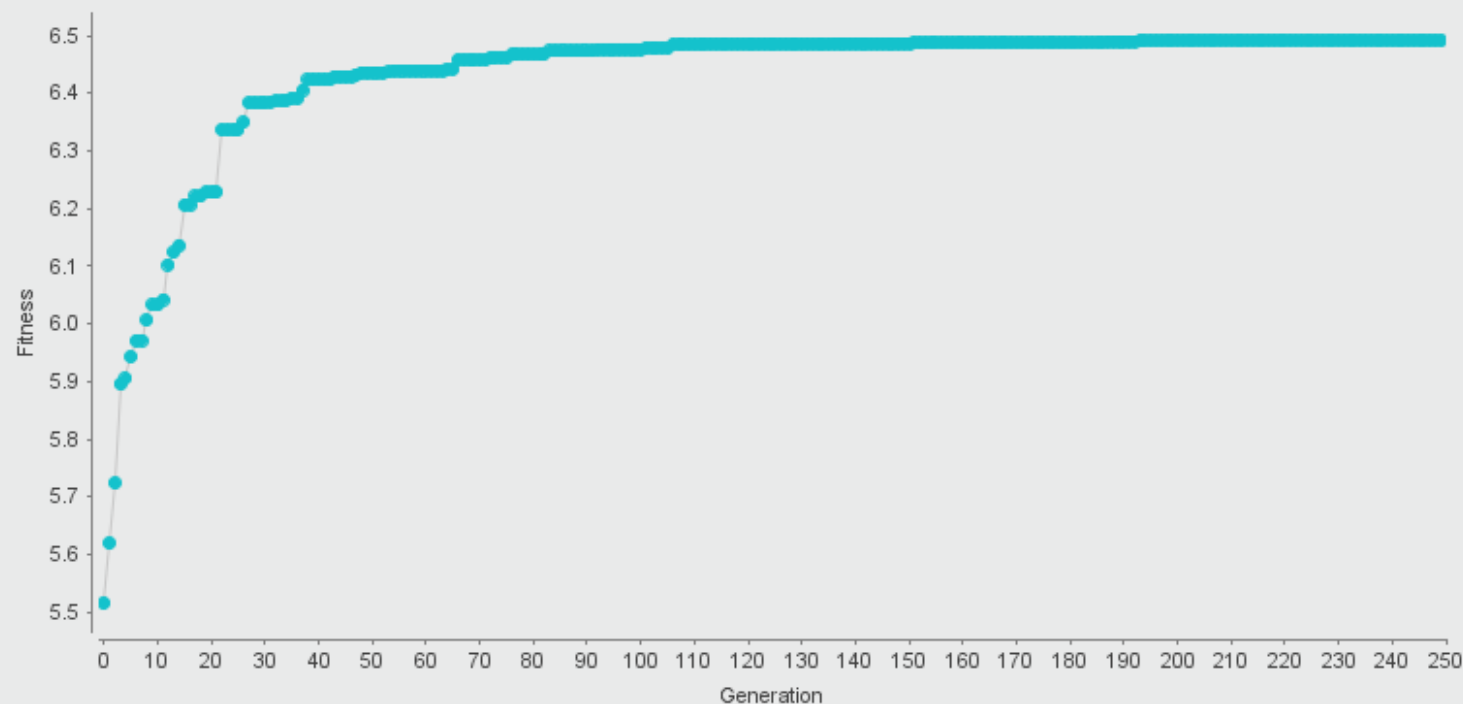
Define Targets

Define Constraints

Optimization Parameters

Running Optimization

The optimization is now running and determines the best input factors to meet your target under the specified constraints. The plot below shows how the fitness of the best solution develops over time. It should converge towards your goal over time.



Optimization done!



Finish

sulphates

chlorides



Views:

Design

Results

Turbo Prep

Auto Model

Deployments

Find data, operators...etc



All Studio

Result History

LinearRegression (Linear Regression)

ModelSimulatorIOObject (Model Simulator)

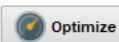
ExampleSet (Multiply)

PerformanceVector (Performance)



Input for Model

chlorides:	<input type="text" value="0.034"/>	ⓘ
alcohol:	<input type="text" value="9"/>	ⓘ
density:	<input type="text" value="0.998"/>	ⓘ
free sulfur dioxide:	<input type="text" value="10"/>	ⓘ
fixed acidity:	<input type="text" value="12.030"/>	ⓘ
total sulfur dioxide:	<input type="text" value="6"/>	ⓘ
volatile acidity:	<input type="text" value="0.170"/>	ⓘ
sulphates:	<input type="text" value="0.994"/>	ⓘ
citric acid:	<input type="text" value="0.001"/>	ⓘ
pH:	<input type="text" value="3.014"/>	ⓘ
residual sugar:	<input type="text" value="5.391"/>	ⓘ



Optimize

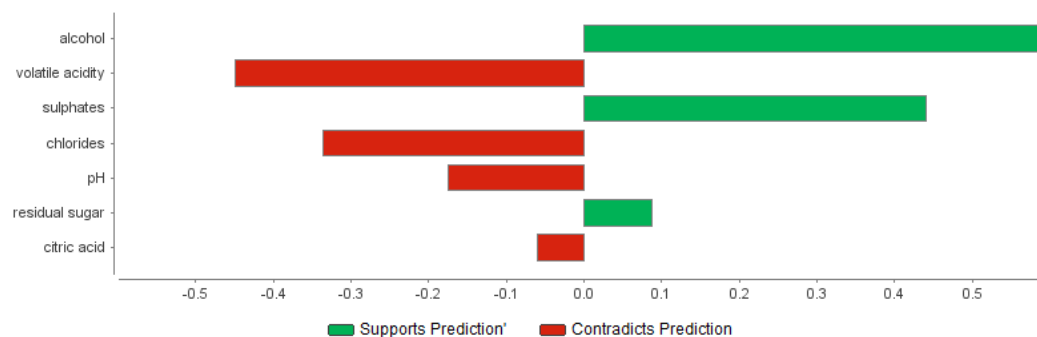
What is [this](#)?

Prediction

Prediction

6.492

Important Factors for Prediction



Interpretation

Select your inputs on the left to see the model's reaction on the right. The prediction of the model is 6.492. The biggest support for this decision is coming from **alcohol**.

Distribution of Predictions

You need to provide a test data set to see a distribution of predictions.

Accuracy

Accuracy can not be calculated: no test data was provided.

Conclusion

- In this lecture, we focused on Linear Regression algorithm.
 - This algorithm derives a linear equation that represents the data
- We also learn also various feature selection algorithms.
- In the next lectures, you will learn optimization in model building with feature extraction algorithms.



QUESTIONS?