Class 10 Lab & HW4 Predictive Modeling Techniques

Learning Objectives:

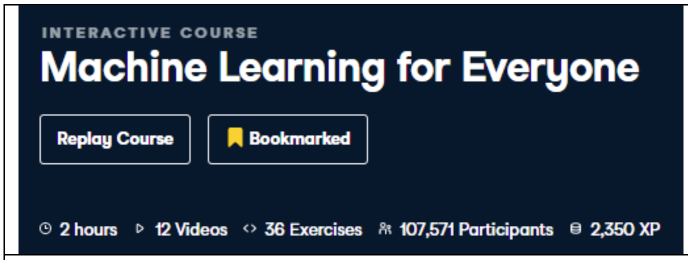
- Understand supervised data mining (DM) methods and process.
- Transform data for different predictive models.
- Perform and compare three popular predictive modeling techniques in RapidMiner.
- Interpret predictive modeling results.

Instructions and Requirements

HW4 includes two parts: Part 1 asks you to complete a theoretical course on DataCamp and Part 2 requires you to complete this week's lab. You are required follow the instruction carefully to perform each step, answer each question in accordance with its required format, and take the required screenshots with date and time. Please submit your answers to the required questions via HW4-Questions Submission on WT Class (under Lessons – Class 10 – HW4 Folder) and compile all your screenshots in HW4-Screenshots.docx and then submit it as an attachment separately via HW4-Screenshot Submission. For HW4-Questions (30 points in total), you have up to two attempts and the one with the higher grade will be counted into your final grade. After your first attempt, you may view your grade on each question (rather than sub-question because many of them are binary answers such as high or low and yes or no) and then figure out why you miss a few questions and then improve your submission. For HW4-Screenshots (30 points in total), please compile the screenshot in Part 1 and five required screenshots with the required format in the template provided on WTCLASS. **If identical screenshots are found, all the involved students will receive zero points and this case will be reported to the dean office.**

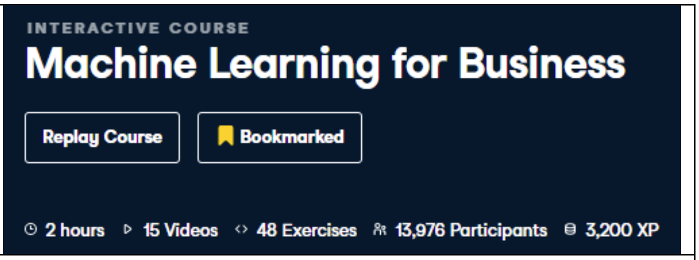
Part 1: Completing An Interactive Course on DataCamp (15 points)

Please choose **one** of the following courses on DataCamp (no coding required). Each course takes you 2 hours and helps you review theoretical parts of analytics.



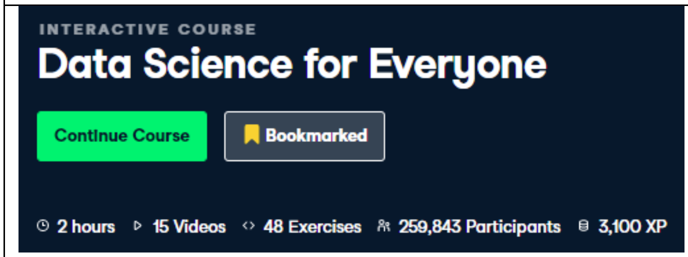
Machine Learning for Everyone

2 hours · 12 Videos · 36 Exercises · 107,571 Participants · 2,350 XP



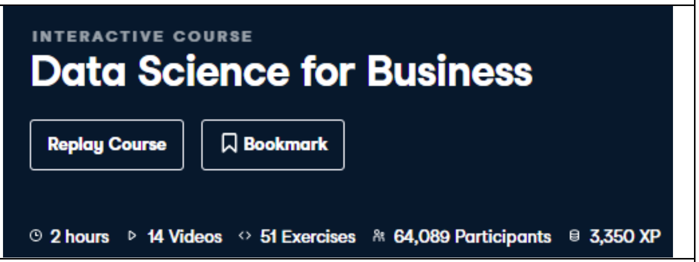
Machine Learning for Business

2 hours · 15 Videos · 48 Exercises · 13,976 Participants · 3,200 XP



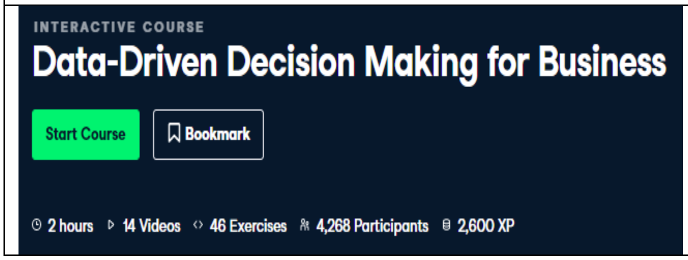
Data Science for Everyone

2 hours · 15 Videos · 48 Exercises · 259,843 Participants · 3,100 XP



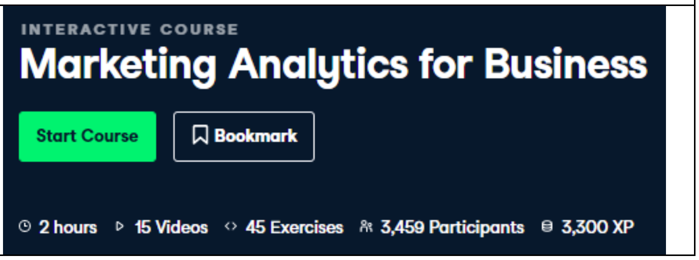
Data Science for Business

2 hours · 14 Videos · 51 Exercises · 64,089 Participants · 3,350 XP



Data-Driven Decision Making for Business

2 hours · 14 Videos · 46 Exercises · 4,268 Participants · 2,600 XP

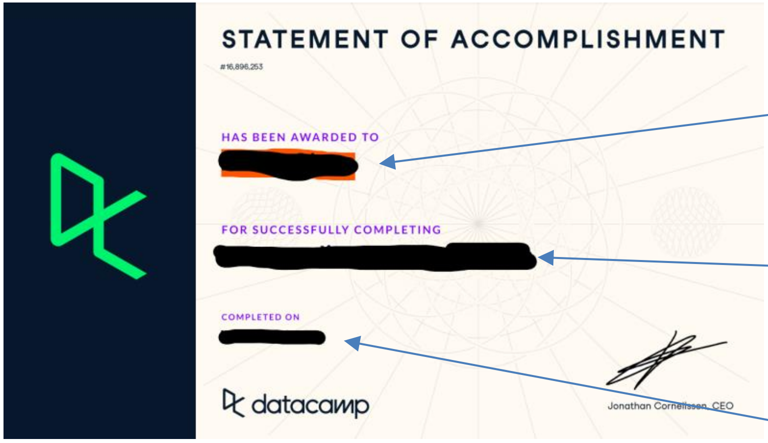


Marketing Analytics for Business

2 hours · 15 Videos · 45 Exercises · 3,459 Participants · 3,300 XP

After completing a course, you will receive **Statement of Accomplishment** (you should receive a copy in your email as well). Please take a screenshot of it to show your completion to earn 15 points (see the requirements below). If you do not know how to take a screenshot, please check this website: https://www.take-a-screenshot.org/.

Requirements:



STATEMENT OF ACCOMPLISHMENT

#16,896,253

HAS BEEN AWARDED TO

[redacted]

FOR SUCCESSFULLY COMPLETING

[redacted]

COMPLETED ON

[redacted]

datacamp

Jonathan Cornelissen, CEO

The name here must be matched with your name on WTClass (if not please explain); otherwise, a zero point will be assigned.

The course name here must be one of the six courses above; otherwise, a zero point will be assigned. Any other course must be approved by the instructor.

The date here must be between March 23rd and April 3rd, 2023; otherwise, a zero point will be assigned.

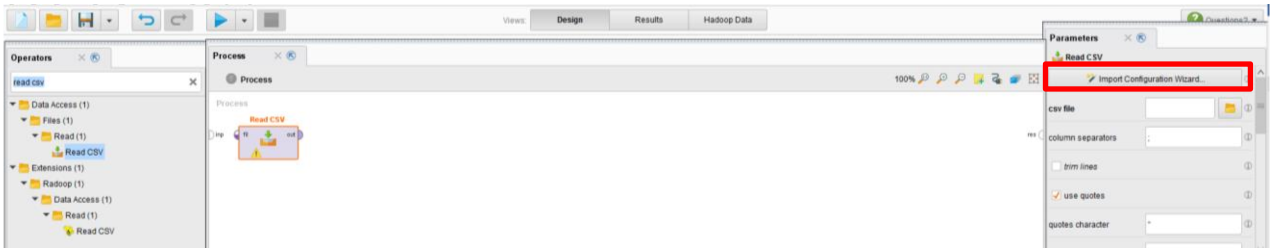
Part 2: Predictive Modeling in RM

Datasets

In this lab session, you are going to practice three supervised DM techniques on RapidMiner: Linear Regression, Decision Tree, and Logistic Regression. Please download the dataset titled "winequality.csv" and “newwines.csv” from WT Class to use with this lab session and homework. The first dataset is a training dataset, which includes chemical characteristics (such as fixed acidity, alcohol level, pH, density, etc.) of 300 wines as well as the quality ratings of these wines graded by experts. Note that attribute values in the dataset are all numerical values. The second dataset is a prediction dataset, which includes the same chemical characteristics of three wines but with quality unknown. We are going to use training dataset (winequality.csv) to build prediction models with different techniques and then use those models to predict the quality of three new wines.

Practice 1 Linear Regression: Modeling and Prediction

1. Add the **Read CSV** operator and then click Click “Import Configuration Wizard” to import your data (please follow the same procedure in Week 09 Lab). After you connect the output port with result port, you can run the process and view your data.



column separators ;

☐ trim lines

☑ use quotes

quotes character "

2. Designate the target attribute (see the screenshot as below)

Different from unsupervised data mining in the previous lab, supervised data mining requires you to assign a target attribute.

2.1 Find the **Set Role** operator under Blending→ Attributes → Names & Roles or search “Set Role” in the search box

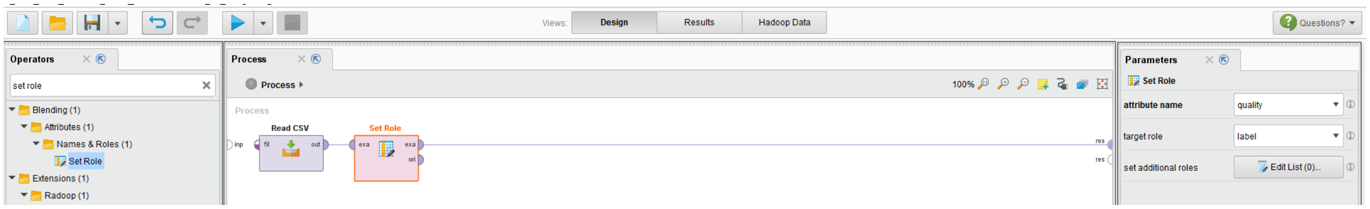
2.2 Add/Drag this operator to your process and set the parameters as below:

2.2.1 Attribute name = quality

2.2.2 Target role = label

2.3 Connect the output of **Read CSV** operator with the input of the **Set Role** operator

2.4 Run this operator by clicking the play icon.



attribute name: quality

target role: label

set additional roles: Edit List (0)...

3. Linear Regression Modeling (see the screenshot as below)

3.1 Find the **Linear Regression** operator under Modeling→ Predictive → Functions or search "Linear Regression" in the search box

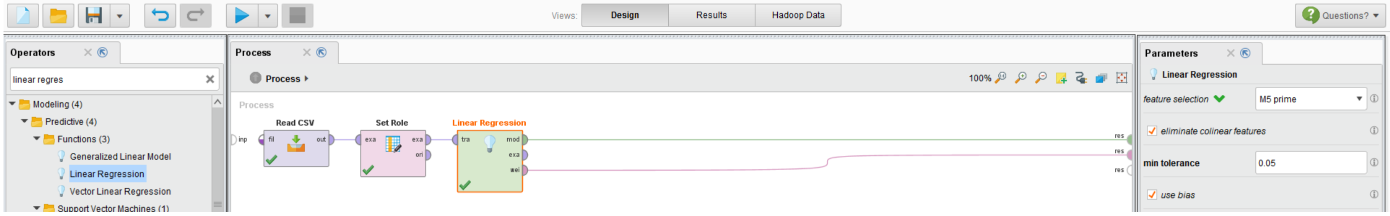
3.2 Add/Drag this operator to your process and use the default parameters (please notice that here we use 0.05 as a cut off value to evaluate the significance of a regression coefficient)

3.3 Connect the output of **Set Role** operator with the input of the **Linear Regression** operator

3.4 Connect the first and third outputs of the **Linear Regression** operator to results:

3.4.1 The first output is model, the regression model is delivered from this output port. This model can now be applied on unseen data sets.

3.4.2 The third output is weights (Attribute Weights) and this port delivers the attribute weights.



feature selection: M5 prime

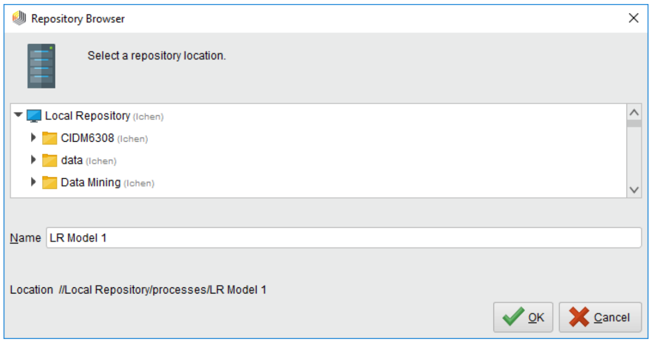
☑ eliminate colinear features

min tolerance: 0.05

☑ use bias

3.5 Save your process as LR model 1 (see the screenshot below) and then run it by clicking the play icon and you will find your results include regression model and attribute weights. Take a screenshot of your regression model (the one with regression coefficients and p values) with date and time (Screenshot 1).

3.6 Answer a few questions as below based on the linear regression model and attribute weights:



Name: LR Model 1

Location: //Local Repository/processes/LR Model 1

3.6.1 Which attribute is a significant and positive predictor of wine quality? Which one is a significant and negative predictor of wine quality?

[Note]: First you need to see whether a regression coefficient is positive or negative. Next, you can use p-value to determine the significance of regression coefficient. When you perform a hypothesis test using statistic methods such as linear regression, a p-value helps you determine the significance of your results. The p-value is a number between 0 and 1 and interpreted in the following way in linear regression (for details about interpreting regression coefficient, please click this link).

- A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis (i.e., H_0: regression coefficient =0), so you reject the null hypothesis, indicating that regression coefficient is not equal to zero at the significant level of 0.05.

- A large p-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.
- P-values very close to the cutoff (0.05) are considered to be marginal (could go either way). Always report the p-value so your readers can draw their own conclusions.

3.6.2 In the attribute weight result, please find the weight of residual sugar (Round it to the third decimal place such as 0.123)?

[Note] Please notice that attribute weight is equal to the regression coefficient in the regression model. In RapidMiner, the default feature selection method is M5 prime (for details, please click this link). Based on this method, some attributes are not included in the regression model. Therefore, their weight is zero.

4. Predicting the quality of new wines

4.1 Similar to Step 1, import the prediction dataset “newwines.csv” using the **Read CSV** operator and then click Click “Import Configuration Wizard”.

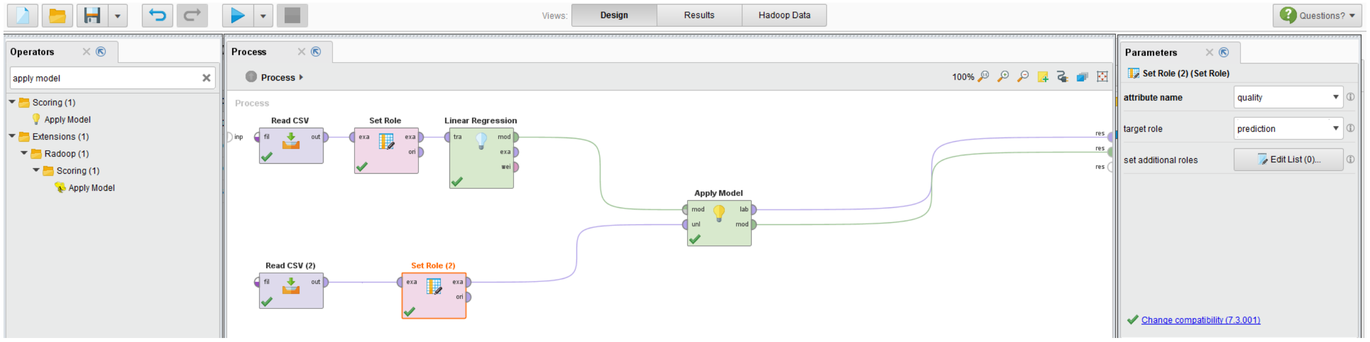
4.2 Add/Drag the **Set Role** operator to your process and set the parameters as below:

4.2.1 Attribute name = quality

4.2.2 Target role = prediction (because you are going to predict the quality of the new wines)

4.3 Add/Drag the **Apply Model** (under Scoring) operator to your process

4.4 Connect the relevant operators in the following way.



Set Role (2) (Set Role)

attribute name: quality

target role: prediction

set additional roles: Edit List (0)...

4.5 Save your process as LR model 2 and run it. Take a screenshot of your prediction results (i.e., the ExampleSet with predicted quality of the five new wines) with date and time (Screenshot 2).

4.6 Looking at the ExampleSet (of prediction dataset) and Linear regression model, answer a few questions as below:

4.6.1 What is the predicted quality of the first new wine (with fixed acidity =5.9)? [round your answer to an integer]

4.6.2 What is the predicted quality of the second new wine (with fixed acidity =9.5)? [round your answer to an integer]

4.6.3 What is the predicted quality of the third new wine (with fixed acidity =5.6)? [round your answer to an integer]

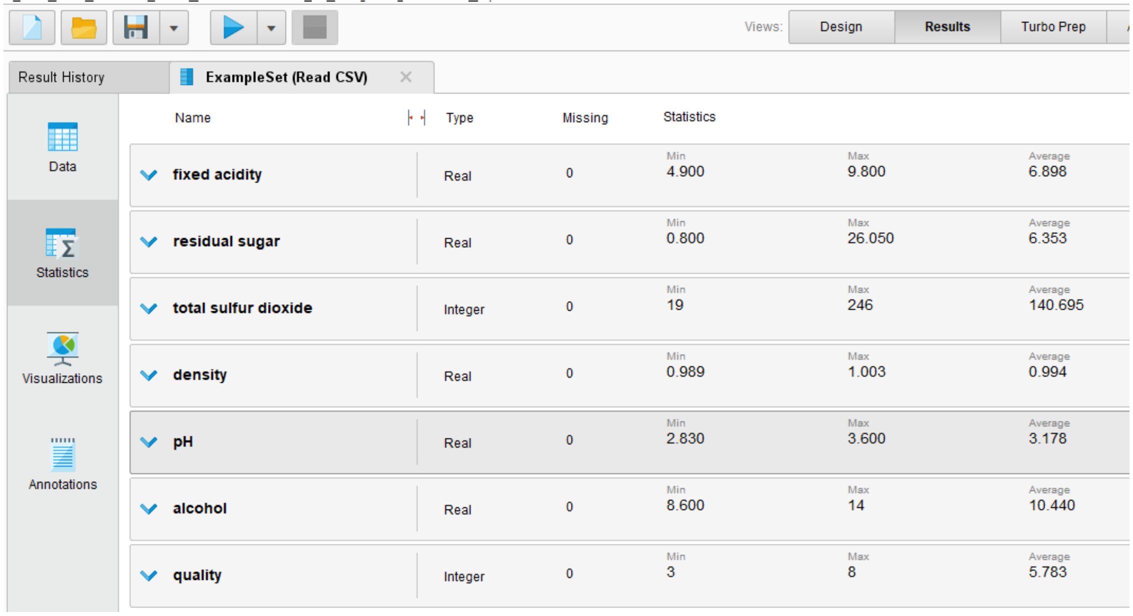
4.6.4 What is the predicted quality of the fourth new wine (with fixed acidity =6.9)? [round your answer to an integer]

4.6.5 What is the predicted quality of the fifth new wine (with fixed acidity =7.9)? [round your answer to an integer]

4.7 Note: In this step, we need to compare the training and prediction datasets. When using linear regression as a predictive model, it is extremely important to remember that the ranges for all attributes in the prediction data must be within the ranges for the corresponding attributes in the training data. This is because a training data set cannot be relied upon to predict a target attrtibute for observations whose values fall outside the training data set's values. If so, you should add the **Filter Example** operator to filter out those examples/records with values outside the range of the training dataset.

In this lab, the prediction dataset only includes five wines, so you can check how many records having at least one attribute outside the range of the corresponding attribute in the training dataset. 0, 1, 2, 3, 4, or 5? Even though you may find such records, we will still keep them for this lab practice.

The stastistics from the 300 wine samples, generated in Step 1, is attached below. Each attribute has a min and max (i.e., the range), which can help you in this question.



Name	Type	Missing	Min	Max	Average
fixed acidity	Real	0	4.900	9.800	6.898
residual sugar	Real	0	0.800	26.050	6.353
total sulfur dioxide	Integer	0	19	246	140.695
density	Real	0	0.989	1.003	0.994
pH	Real	0	2.830	3.600	3.178
alcohol	Real	0	8.600	14	10.440
quality	Integer	0	3	8	5.783

For example, for the first wine, fixed acidity =5.9, which is between 4.9 and 9.8 (see the screenshot above); residual sugar=4.5, which is between 0.8 and 26.05; total sulfur dioxide=143, which is between 19 and 246; density=0.9912, which is between 0.989 and 1.003; pH=2.96, which is between 2.83 and 3.60; alcohol=13.8, which is between 8.6 and 14. Therefore, each attribute of the first wine is within the range of their corresponding attribute in the training dataset.

fixed acidity	residual sugar	total sulfur dioxide	density	pH	alcohol
5.9	4.5	143	0.9912	2.96	13.8

Practice 2 Decision Tree: Modeling and Prediction

5. Data preparation (see the screenshot as below)

5.1 We are going to start with opening the same dataset "winequality.csv" in Excel and then make a few changes.

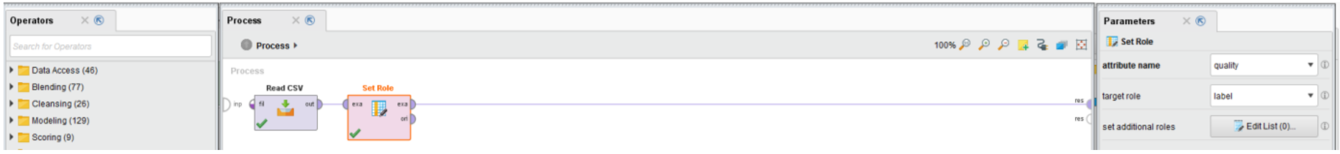
5.2 Because decision trees deal with categorical target attributes, we need to change the data type of the target attribute, quality, from numerical to categorical in Excel based on the following method:

- 3, 4, and 5 → low
- 6, 7, and 8 → high

5.3 Save the dataset as a new CSV file "winequality2.csv".

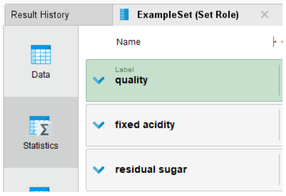
5.4 Import the data using the Read CSV operator and follow the same procedure in Step 1.

5.5 Designate the target attribute using the Set Role operator and follow the same procedure in Step 2.



attribute name: quality
target role: label
set additional roles: Edit List (0)...

5.6 Run the process and in your ExampleSet click the Statistics view at the left sidebar (see the screenshot in the right). Review the results there and answer the question: among 300 observations, how many are rated as high quality?

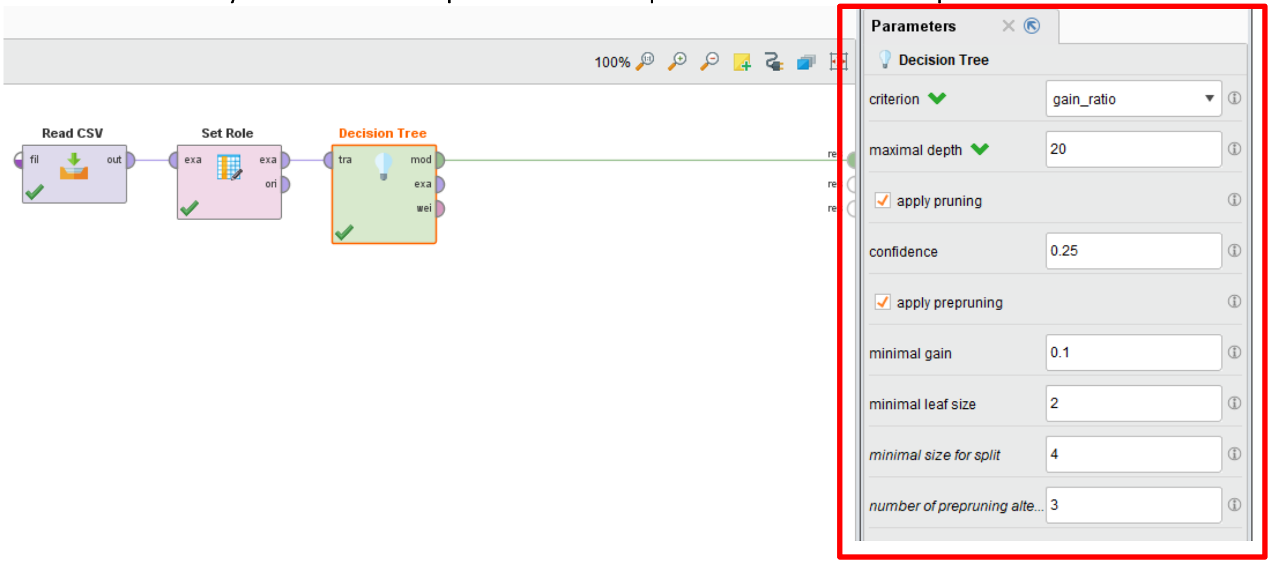


6. Build Decision Tree Model (see the screenshot as below)

6.1 Find the **Decision Tree** operator under Modeling→ Predictive → Trees or search "Decision Tree" in the search box.

6.2 Add/Drag this operator to your process and **use the parameters as below (see the parameters in the red box).**

6.3 Make sure you connect this operator with the previous one and result port.

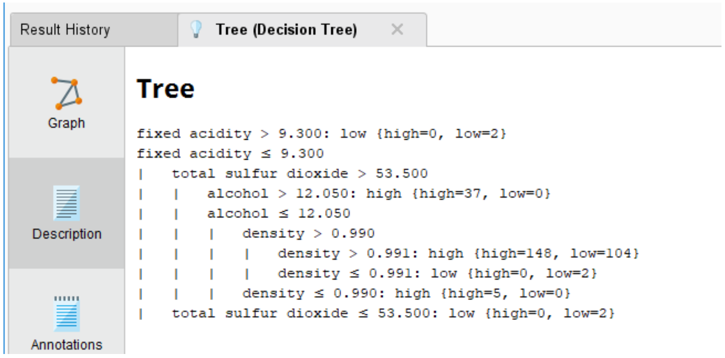


Parameters

Decision Tree

criterion	gain_ratio
maximal depth	20
☑ apply pruning	
confidence	0.25
☑ apply prepruning	
minimal gain	0.1
minimal leaf size	2
minimal size for split	4
number of prepruning alte...	3

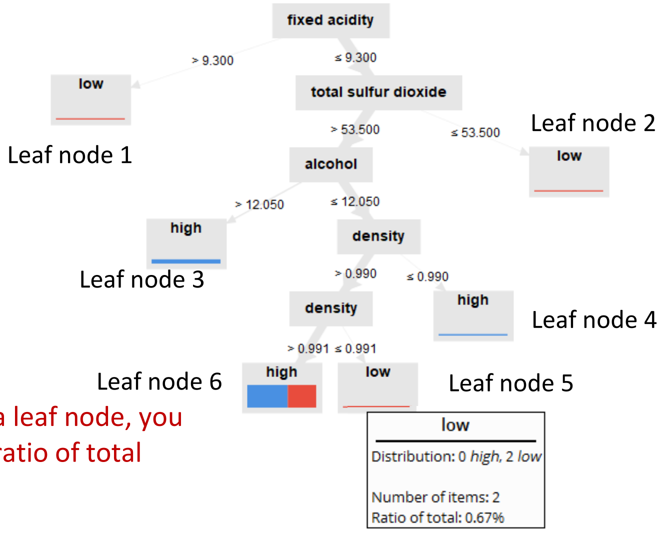
6.4 Save your process as DTmodel1 and then run this process and you will see a decision tree model and description (see the two screenshots below). Take a screenshot of your decision tree model (the graph one) with date and time (Screenshot 3).



Tree

```
fixed acidity > 9.300: low {high=0, low=2}
fixed acidity ≤ 9.300
|   total sulfur dioxide > 53.500
|   |   alcohol > 12.050: high {high=37, low=0}
|   |   alcohol ≤ 12.050
|   |   |   density > 0.990
|   |   |   |   density > 0.991: high {high=148, low=104}
|   |   |   |   density ≤ 0.991: low {high=0, low=2}
|   |   |   density ≤ 0.990: high {high=5, low=0}
|   total sulfur dioxide ≤ 53.500: low {high=0, low=2}
```

The description of the decion tree



low

Distribution: 0 *high*, 2 *low*

Number of items: 2

Ratio of total: 0.67%

If you move your cursor to a leaf node, you will see the class, size, and ratio of total

Note: your tree may look different from the one below. You can zoom in or out this tree and move it as well. Similar to linear regression model, the decision tree model does not include all attributes in your dataset due to feature selection and pruning.

6.5 Observe this decision tree and read the description of the tree above and then answer the following questions:

6.5.1 Excluding the root node, how many split nodes in this tree?

6.5.2 What is the size of the leaf node 3?

6.5.3 Among the six leaf nodes, how many generate a pure class (i.e, single color in the leaf node)?

6.5.4 Please use this decision tree to determine the quality of the following two wines (from the top to bottom):

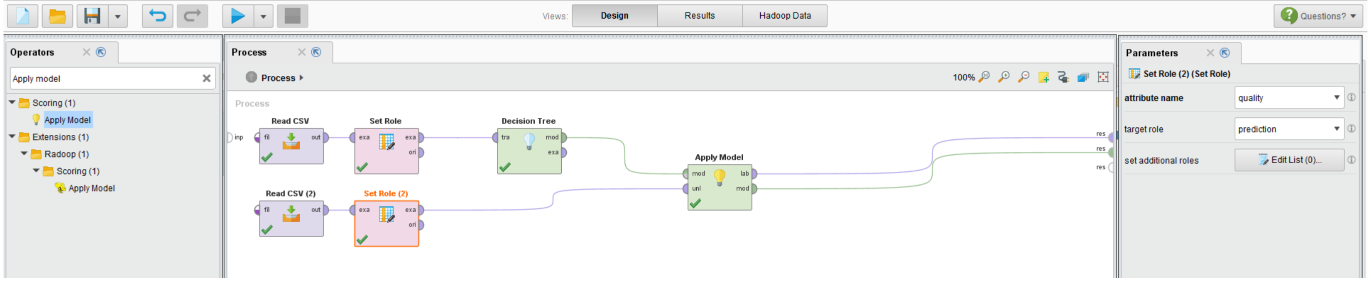
fixed acidity	residual sugar	total sulfur dioxide	density	pH	alcohol	Quality (high or low)
9.14	2.9	102	0.9912	3.17	11	
7	6.4	50	0.9954	3.13	9.5	

7. Apply the Decision Tree Model (see the screenshot as below)

7.1 Follow the same procedure in Step 4.1-4.4.

7.2 Connect the relevant operators in the following way.

7.3 Save the process as DTmodel2 and run this process.



Set Role (2) (Set Role)

attribute name: quality

target role: prediction

set additional roles: Edit List (0)...

7.4 Save the process as DTmodel2 and run this process. Take a screenshot of your prediction results with date and time (Screenshot 4).

7.5 Looking at the ExampleSet (of prediction dataset), answer a few questions as below:

7.5.1 What is the predicted quality of the first new wine (with fixed acidity =5.9)? [low or high]

7.5.2 What is the predicted quality of the second new wine (with fixed acidity =9.5)? [low or high]

7.5.3 What is the predicted quality of the third new wine (with fixed acidity =5.6)? [low or high]

7.5.4 What is the predicted quality of the fourth new wine (with fixed acidity =6.9)? [low or high]

7.5.5 What is the predicted quality of the fifth new wine (with fixed acidity =7.9)? [low or high]

Practice 3 Logistic Regression: Modeling and Prediction

8. Data preparation

8.1 We are going to start with the same dataset “winequality2.csv” in Step 5, with one change below.

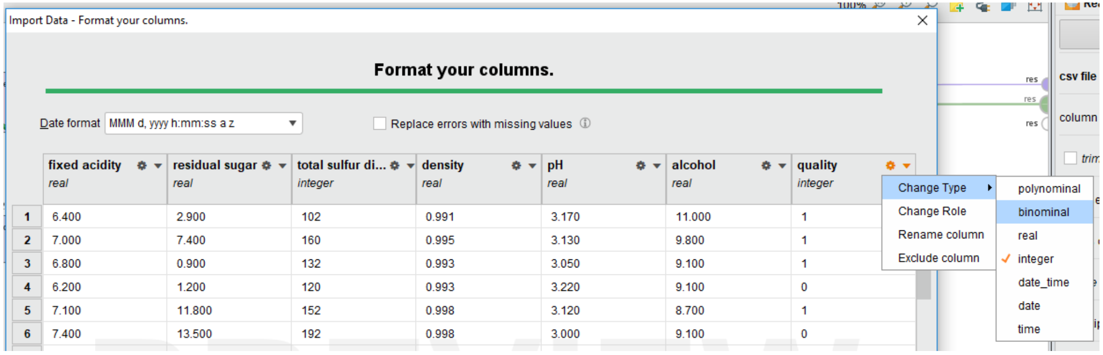
8.2 Because logistic regression deals with binary numerical target attributes (i.e., 0, 1), we need to change the data type of the target attribute, quality, from categorical to binary or binomial (0-1) in Excel based on the following method:

- code low → 0
- code high→ 1

Attention: Recent versions of RM can handle binary or binomial target attribute directly, either low-high, or 0-1. Therefore, you can run logistic regression without such a transformation in RM. However, for this practice, let’s still perform such a transformation because some other software packages they you may be using in the future may require such as transformation.

8.3 Save the dataset as a new CSV file “winequality3.csv”.

8.4 Import the data using the Read CSV operator using the same procedure before; in Step 3 of Data import wizard – Format your columns, you need to change data type of quality to binominal (see the screenshot as below) and then click Finish.



Import Data - Format your columns.

Format your columns.

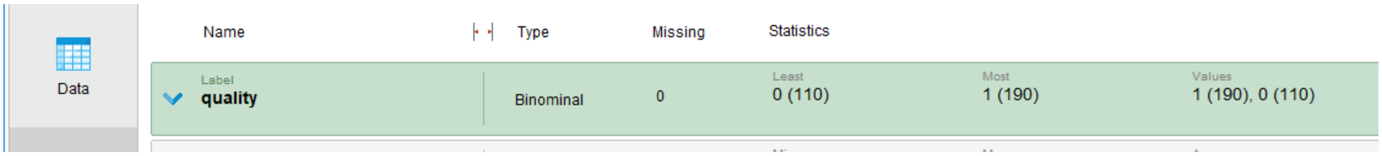
Date format MMM d, yyyy h:mm:ss a z

☐ Replace errors with missing values

	fixed acidity real	residual sugar real	total sulfur di... integer	density real	pH real	alcohol real	quality integer
1	6.400	2.900	102	0.991	3.170	11.000	1
2	7.000	7.400	160	0.995	3.130	9.800	1
3	6.800	0.900	132	0.993	3.050	9.100	1
4	6.200	1.200	120	0.993	3.220	9.100	0
5	7.100	11.800	152	0.998	3.120	8.700	1
6	7.400	13.500	192	0.998	3.000	9.100	0

8.5 Designate the target attribute using the Set Role operator and follow the same procedure in Step 2.

8.6 Run the process and look at Statistics in your ExampleSet. You can see the data type of quality is Binominal and the number of 1s and 0s.



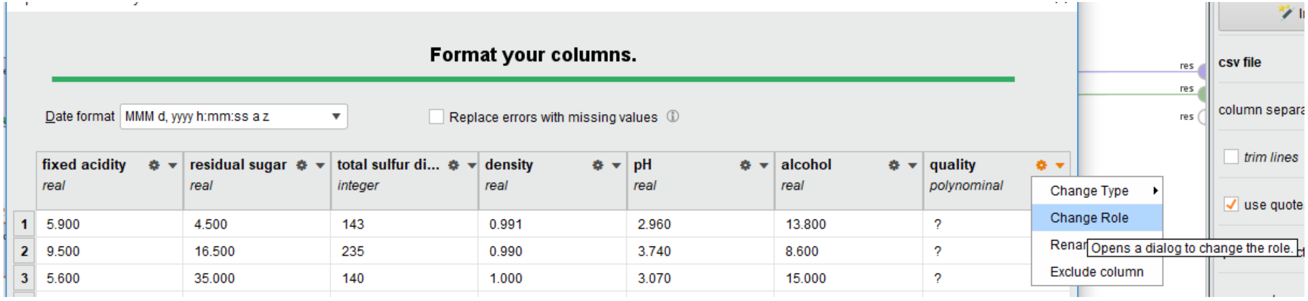
Name	Type	Missing	Statistics		
Label quality	Binominal	0	Least 0 (110)	Most 1 (190)	Values 1 (190), 0 (110)

9. Build and Apply Logistic regression Model

9.1 Find the **Logistic Regression** operator under Modeling→ Predictive → Logistic Regression or search "Logistic Regression" in the search box.

9.2 Add/Drag this operator to your process and use the default parameters.

9.3 Import your prediction dataset using the **Read CSV** operator and this time we are going to use an alternative method to avoid using the Set Role operator again in your process for the prediction dataset: in Step 3-Format your columns, please change the role of quality from attribute to prediction and change the data type from polynominal to binominal (see the screenshots as below).



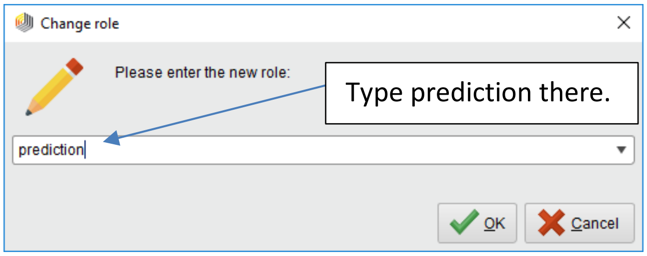
Format your columns.

Date format MMM d, yyyy h:mm:ss a z

☐ Replace errors with missing values

	fixed acidity real	residual sugar real	total sulfur di... integer	density real	pH real	alcohol real	quality polynominal
1	5.900	4.500	143	0.991	2.960	13.800	?
2	9.500	16.500	235	0.990	3.740	8.600	?
3	5.600	35.000	140	1.000	3.070	15.000	?

Opens a dialog to change the role.

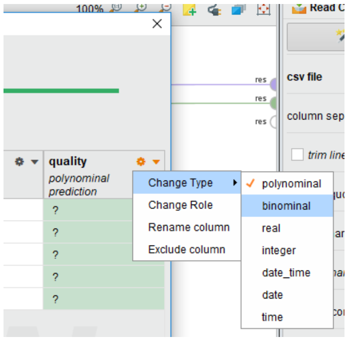


Change role

Please enter the new role:

prediction

Type prediction there.

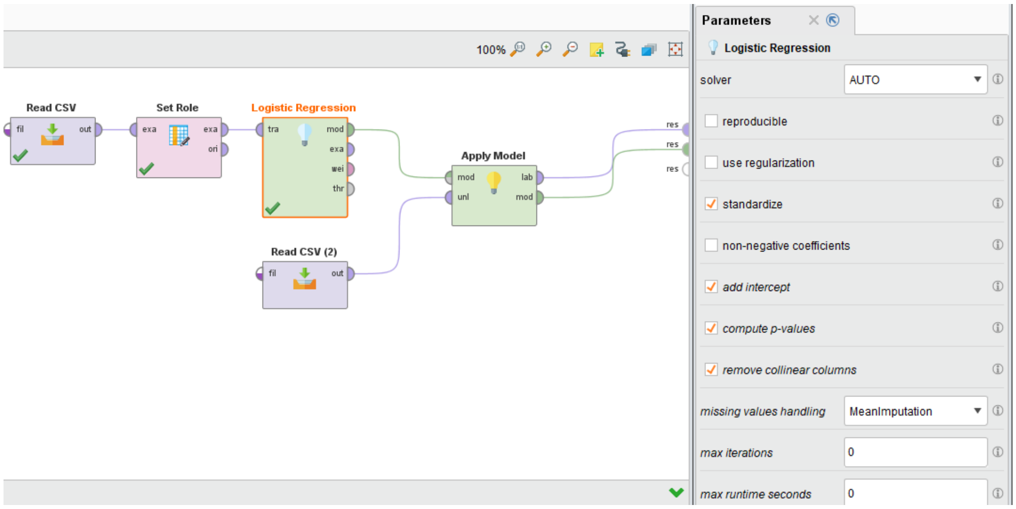


quality
polynominal
prediction

?
?
?
?
?

9.4 Add/drag the **Apply Model** operator use the default settings.

9.5 Connect the relevant operators in the following way and save your process as LogRmodel.



Parameters

Logistic Regression

solver: AUTO

☐ reproducible

☐ use regularization

☑ standardize

☐ non-negative coefficients

☑ *add intercept*

☑ *compute p-values*

☑ *remove collinear columns*

missing values handling: MeanImputation

max iterations: 0

max runtime seconds: 0

9.6 Run the process and take a screenshot of your prediction results with date and time (Screenshot 5). Please answer the following questions:

9.6.1 What is the predicted quality of the first new wine (with fixed acidity =5.9)? [1 or 0]
9.6.2 What is the predicted quality of the second new wine (with fixed acidity =9.5)? [1 or 0]
9.6.3 What is the predicted quality of the third new wine (with fixed acidity =5.6)? [1 or 0]
9.6.4 What is the predicted quality of the fourth new wine (with fixed acidity =6.9)? [1 or 0]
9.6.5 What is the predicted quality of the fifth new wine (with fixed acidity =7.9)? [1 or 0]

10. Model Comparisons

10.1 Typically, different prediction methods may generate different predictions. Compare your results in Step 4.6, Step 7.5, and Step 9.6 to see whether or not they generate the same conclusion. Here, in order to see whether or not the three methods generate the same predictions, you need to convert the predicted quality in Step 4.6 to High/Low using the same algorithm in Step 5.2. Likewise, you need to convert the predicted quality in Step 9.6 to High/Low using the reversed algorithm in Step 8.2.

10.2 For each new wine, please indicate whether or not the three methods generate the same prediction results in the following table.

	Linear Regression	Decision Tree	Logistic Regression	Same prediction?
The predicted quality of the wine with fixed acidity =5.9				Yes or No
The predicted quality of the wine with fixed acidity =9.5				Yes or No
The predicted quality of the wine with fixed acidity =5.6				Yes or No
The predicted quality of the wine with fixed acidity =6.9				Yes or No
The predicted quality of the wine with fixed acidity =7.9				Yes or No

Appendices FAQs

Q: I don't have a huge background on Business Statistics. I read through the slides in the "Optional Material" folder and still can't get a grasp on this information to answer the below question. Which attribute is a significant and positive predictor of wine quality? Which one is a significant and negative predictor of wine quality?

A: As explained in our lab instruction, first you need to see whether a regression coefficient is positive or negative. Next, you can use p-value to determine the significance of regression coefficient. When you perform a hypothesis test using statistic methods such as linear regression, a p-value helps you determine the significance of your results. The p-value is a number between 0 and 1. The smaller p value, the more significant relationship is (lower chance to be zero).

I use the paper I presented in the conference last weekend as an example to explain this for you. First of all, see the regression coefficient (Beta in the following table). Is it smaller than 0 (negative) or greater than 0 (positive)? In the following case, the regression coefficient of publication year is positive. Next, please see the p value. If p value is smaller than 0.05, the regression coefficient is significant at 0.05 level. The p value for publication year is 0.00, so it is significant. Overall, publication year is a positive and significant factor.

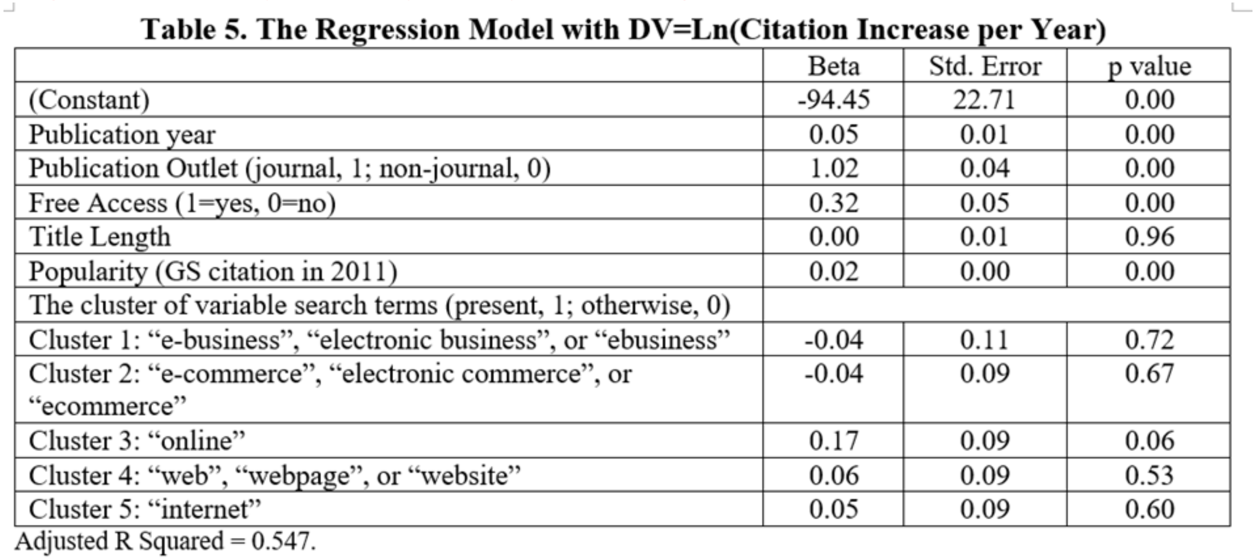


Table 5. The Regression Model with DV=Ln(Citation Increase per Year)

	Beta	Std. Error	p value
(Constant)	-94.45	22.71	0.00
Publication year	0.05	0.01	0.00
Publication Outlet (journal, 1; non-journal, 0)	1.02	0.04	0.00
Free Access (1=yes, 0=no)	0.32	0.05	0.00
Title Length	0.00	0.01	0.96
Popularity (GS citation in 2011)	0.02	0.00	0.00
The cluster of variable search terms (present, 1; otherwise, 0)			
Cluster 1: “e-business”, “electronic business”, or “ebusiness”	-0.04	0.11	0.72
Cluster 2: “e-commerce”, “electronic commerce”, or “ecommerce”	-0.04	0.09	0.67
Cluster 3: “online”	0.17	0.09	0.06
Cluster 4: “web”, “webpage”, or “website”	0.06	0.09	0.53
Cluster 5: “internet”	0.05	0.09	0.60

Adjusted R Squared = 0.547.

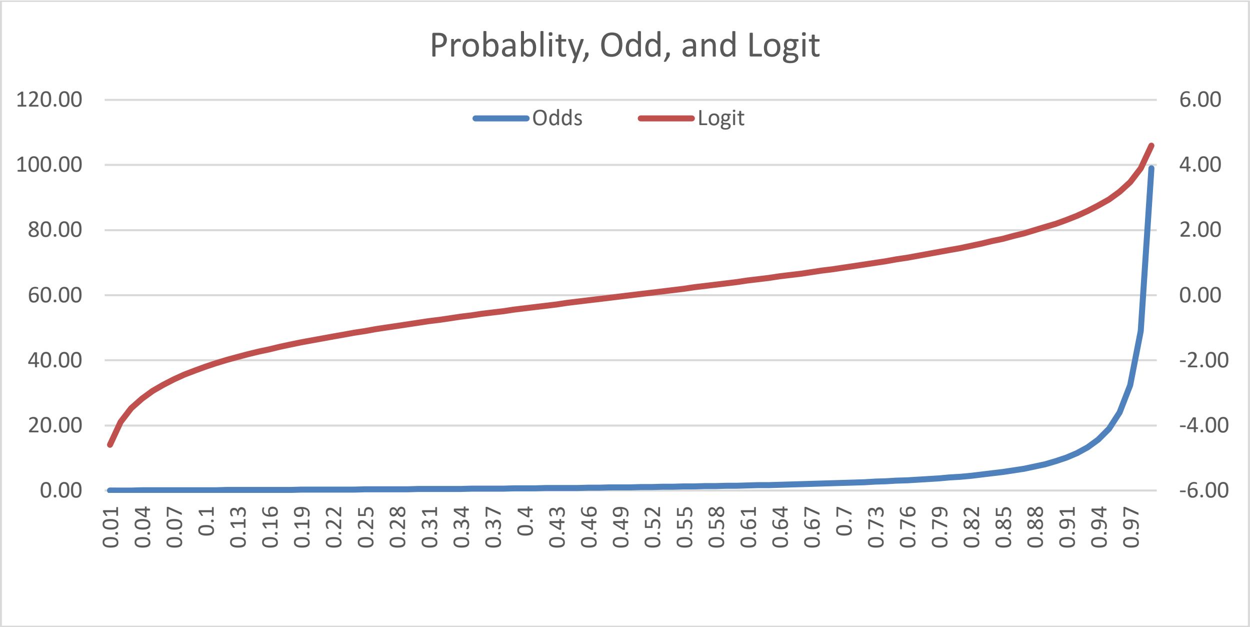
Q: I am still confused about how to compute odds and log odds. Could you help me with that?

A: As shown in the example in Class 10 Lecture, you either get an offer or not. If the probability to get a job offer is =0.8, then the probability of not getting an offer 1-p=0.2, and thus the odds ratio of getting a job is 0.8/0.2=4 and its log odds is ln(4)=1.39. If you are interested in not getting a job, the odds ratio of not getting a job is 0.2/0.8=0.25 and its log odds is ln(0.25)=-1.39.

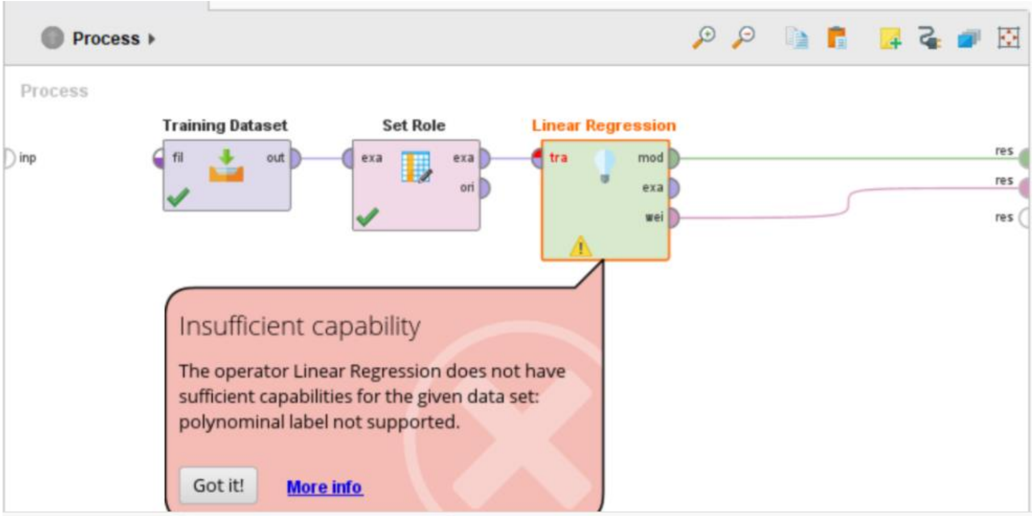
I use the following table to help you get a better idea:

Probability of getting an offer	Odds ratio	Log odds	Probability of not getting an offer	Odds ratio	Log odds
0.1	0.11	-2.20	0.9	9.00	2.20
0.3	0.43	-0.85	0.7	2.33	0.85
0.5	1.00	0.00	0.5	1.00	0.00
0.7	2.33	0.85	0.3	0.43	-0.85
0.9	9.00	2.20	0.1	0.11	-2.20

As you see, the probability originally ranges from 0 to 1. Its odds ratio ranges from 0 to positive infinity (+∞) and its logit ranges from -∞ to +∞, as shown in the following chart.



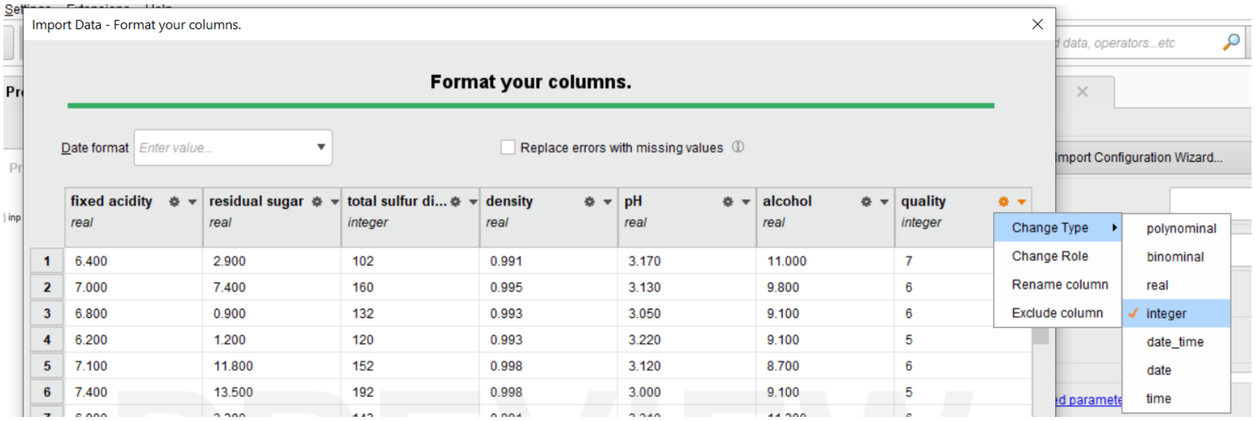
Q: Hi Dr. Chen, I got the following error when running my linear regression model. Could you help me with that? Thank you!



Insufficient capability

The operator Linear Regression does not have sufficient capabilities for the given data set: polynominal label not supported.

A: When you import the dataset to RM via Read CSV operator, quality is supposed to be integer. Your error message indicates that quality is a polynominal in your dataset. If you so, please change it type to integer by following the screenshot below.


Import Data - Format your columns.

Format your columns.

Date format: Enter value...

☐ Replace errors with missing values

	fixed acidity real	residual sugar real	total sulfur di... integer	density real	pH real	alcohol real	quality integer
1	6.400	2.900	102	0.991	3.170	11.000	7
2	7.000	7.400	160	0.995	3.130	9.800	6
3	6.800	0.900	132	0.993	3.050	9.100	6
4	6.200	1.200	120	0.993	3.220	9.100	5
5	7.100	11.800	152	0.998	3.120	8.700	6
6	7.400	13.500	192	0.998	3.000	9.100	5