Data Mining Methods Week 5 RapidMiner Lab: Classification using Naive Bayes & Logistic Regression

Learning Objectives:

- **Understand how overfitting can be solved by pruning (esp. Pre-pruning)**
- **Understand and apply Naïve Bayes classification technique**
- **Grasp condition probability in Naïve Bayes classification**
- **Understand and apply Logistic Regression technique**

Datasets: Please download the datasets titled "vote-train.csv" and "vote-predict-vote.csv" from WT class for this lab session. The training dataset shows the votes of the U.S. House of Representatives Congressmen on the 16 key votes in 1987 as well as their party affiliation. A Congressmen can either be a Democrat or a Republican. Votes can either be Yes, No or Abstain. The other dataset is in the same structure, but includes some values that we'd like to predict.

Requirements: You are not required to take screenshots for this lab. Please submit your answers via HW2 Submission.

1. Data Preparation

1.1 Import the "vote-train.csv" dataset into RapidMiner (RM) using the Read CSV operator with the default settings in the Import Configuration Wizard. Run this operator and check out the results.

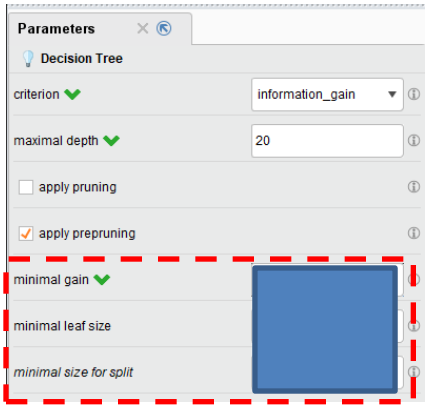
1.2 Now, assume that you are a journalist, who is trying to predict whether the "education-spending" Bill introduced to the Congress will pass or not. Let's assume that this Bill will pass if there are **172 Yes votes** for it. Currently, you have the training dataset "vote-train.csv" with actual vote results of 400 Congressmen (Hint: you can check out the Statistics View – Values column to see how many votes are Yes for this Bill). In the "vote-predict-vote.csv" dataset, you will see 35 observations with missing values for the "education-spending" attribute (These people have not voted yet). You need to predict these 35 observations' votes based on a classification model built on the training dataset.

2 Modeling and Pruning – Decision Tree

2.1 First, build a learning (i.e., training) model using the training dataset and the Decision Tree operator. This is similar to what we learned in Week 4. Please remember to configure the Set Role operator.

2.2 We are going to perform an experiment by building six DT models by specifying the following configurations/parameters:

Model	Minimal gain	Minimal leaf size	Minimal size for split	Other Parameters
Model 1	0.1	3	10	Criterion= information gain
Model 2	0.1	3	50	Maximal depth =20
Model 3	0.1	10	50	Uncheck “apply pruning”
Model 4	0.05	5	10	Check “apply prepruning”
Model 5	0.05	5	50	Number of prepruning
Model 6	0.05	10	50	alternatives =10



Tip 1: Sometimes your new settings may not be applied; if so, please press Enter in your keyboard after making those change in the setting.

Tip 2: Check ⓘ to see what each parameter means (for details about all the parameters, output, or input about DT, please this link).

Tip 3: In order to help you compare those models, you can either view all of them in Result History or save each model’s tree graph in a MS WORD file.

2.3 Run the six models and then answer the following questions:

2.3.1 Please count how many split nodes (including the root node) and leaf nodes each tree graph has.

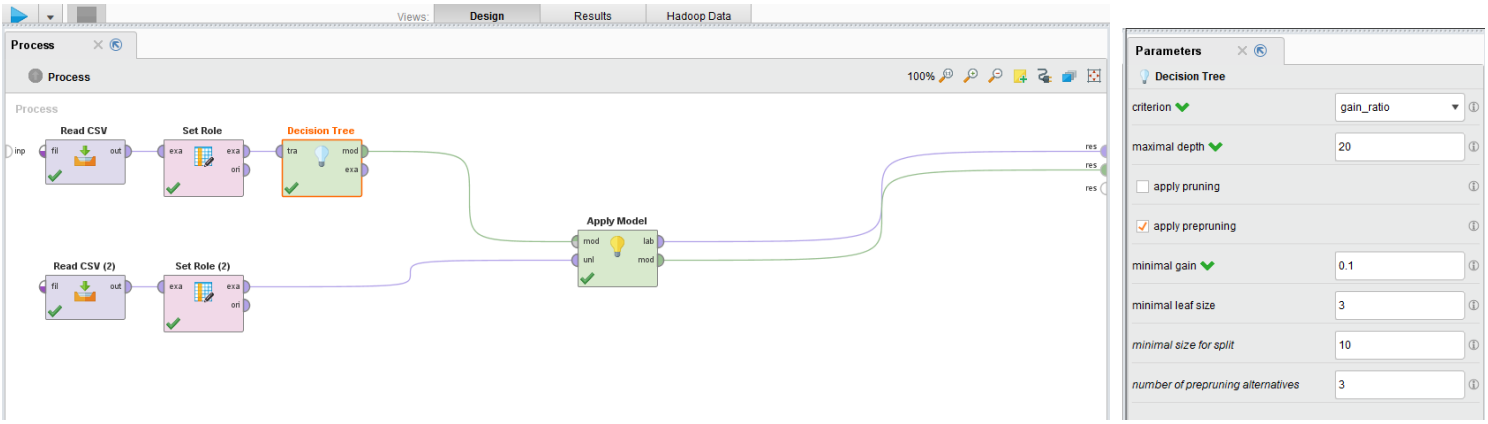
Model	The number of split nodes (including the root node)	The number of leaf nodes
Model 1		
Model 2		
Model 3		
Model 4		
Model 5		
Model 6		

2.3.2 Think about the question: Which model is most likely to have the overfitting problem? Why?

Even though you do not need to type your answers in HW2 Submission for this experiment, I hope you have an idea about how Minimal gain, minimal size for split, and minimal leaf size may change the shape of your decision tree or solve the over-fitting issue.

2.4 Next, import the 35 observations into your process as well ("vote-predict-vote.csv") and set the role of the "education-spending" to prediction.

2.5 Combine these two streams on an "Apply Model" operator and observe the prediction results for 35 observations. Below is the entire process. Please use the following configuration of decision tree (Criterion=gain_ratio, maximal depth = 20, uncheck “apply pruning”, minimal gain =0.1, minimal leaf size= 3; minimal size for split= 10; Number of prepruning alternatives =3).

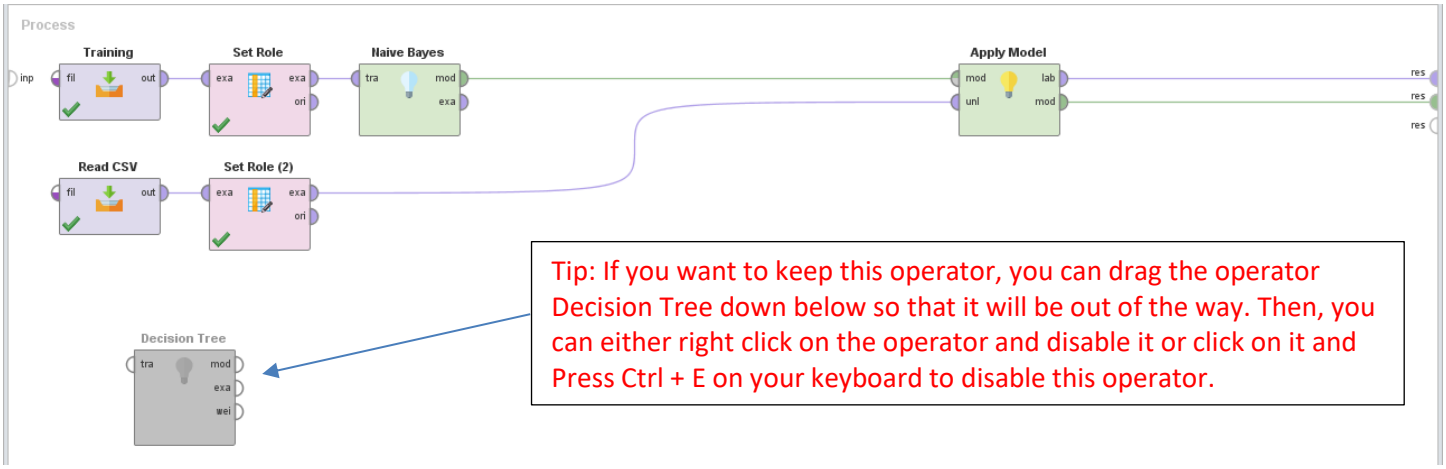


2.6 Please take a look at the statistics view of the ExampleSet and then answer the questions: How many people are predicted to vote “y” (i.e., Yes)? Type a whole number here. Would the Bill pass the Congress? Type Yes or No here.

3 Modeling – Naïve Bayes

3.1 Now, please do the same thing as Step 2 above, but this time we use the Naive Bayes operator, instead of the Decision tree operator. Use the default configuration for the Naive Bayes operator (check laplace correction).

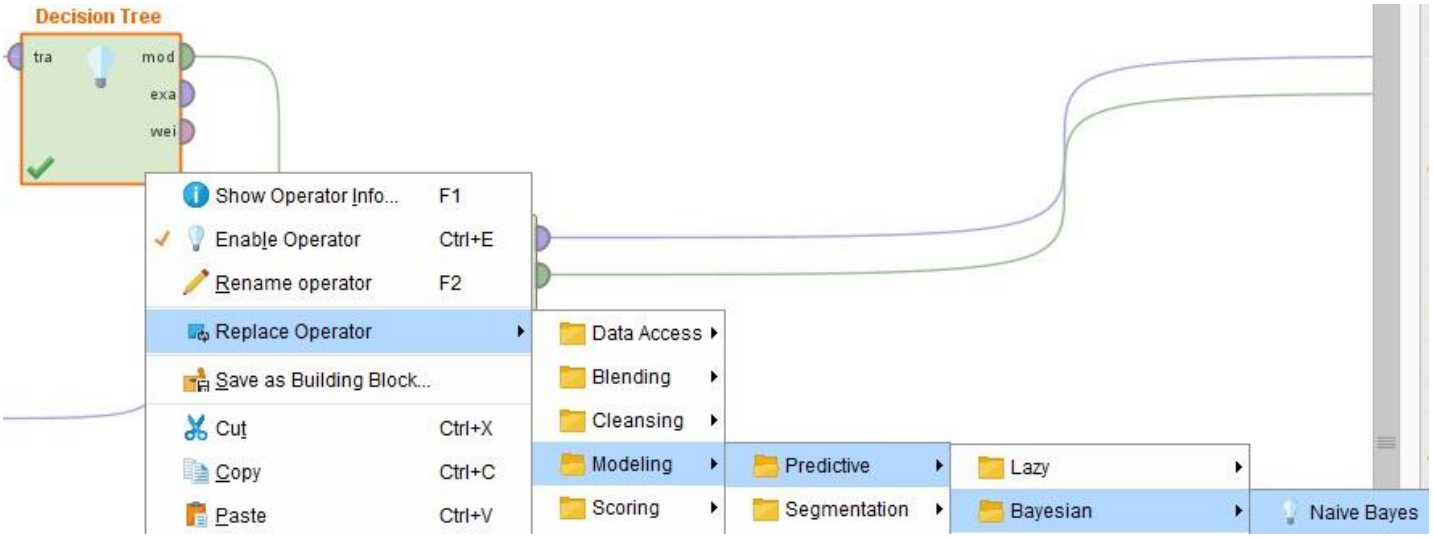
3.2 Run your RM process.



3.3 Take a look at the Statistic view of your prediction results and then answer the questions: How many people are predicted to vote "y" (i.e., Yes)? Type a whole number here. Would the Bill pass the Congress? Type Yes or No here.

Note: Typically, we use two or more classification methods in order to confirm a predicted outcome. Such a doing is called **triangulation**. Next week, we are going to conduct model comparison.

Tip: You can also just replace the Decision Tree operator by Naïve Bayes directly (as shown below).

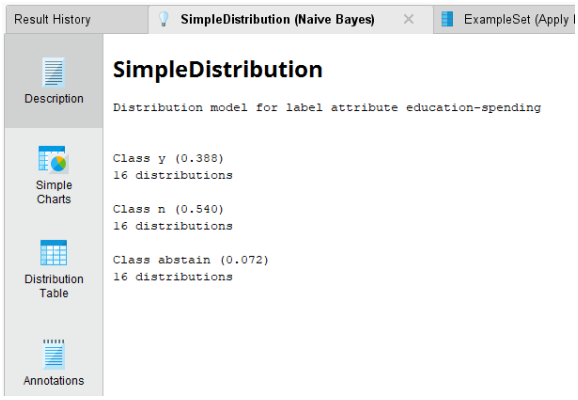


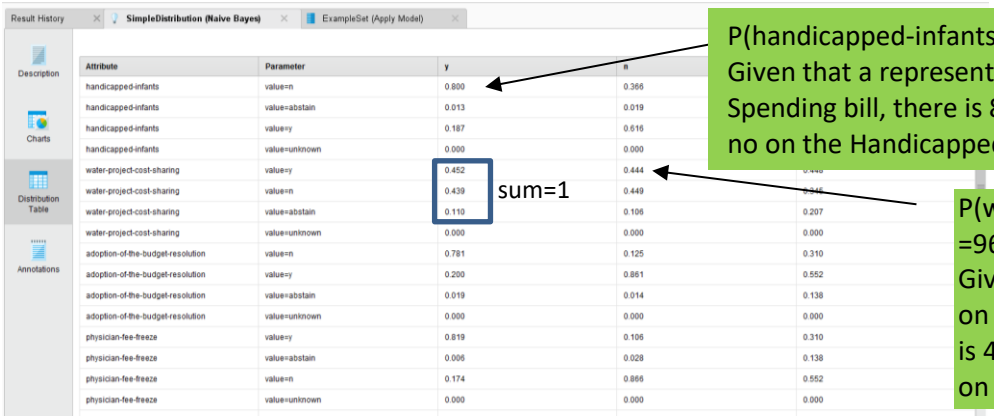
4 Understand Naïve Bayes Algorithm and Result

The "SimpleDistribution" can help us understand how Naïve Bayes works. Under the Description view, you can find that the probability of each vote decision on Educational Spending (ES) in the training dataset:

P(ES=y)=0.388, P(ES=n)=0.540, P(ES=abstain)=0.072.

Under the Distribution Table, we can view all conditional probabilities given that ES=y, n, or abstain. All the probabilities are displayed by the order of all the attributes under the Charts view (Please ignore value=Unknown). The following chart shows you what each number means.





P(handicapped-infants=n|ES=y)=124/155=0.800
Given that a representative votes Yes on the Educational Spending bill, there is 80.0% probability that s/he votes no on the Handicapped Infants bill.

P(water-project-cost-sharing=y|ES=n) =96/216=0.444
Given that a representative votes No on the Educational Spending bill, there is 44.4% probability that s/he votes Yes on the Water Project Cost Sharing bill.

(In order to view other details about the Distribution Table, please check the documentation Naive Bayes on RM)

4.1 Please find the following conditional probabilities directly from the Distribution Table. Since by default RM rounds those probabilities to the third decimal place, you are required to round your answers to the third decimal place such 0.345 too.

- P(immigration=n |ES=y)= _____
- P(Party= republican |ES=y)= _____
- P(Party=democrat |ES=n)= _____
- P(crime=n | ES=n)= _____
- P(anti-satellite-test-ban=y |ES=abstain)= _____

4.2: When a representative votes Yes on the Educational Spending bill, what are the top three possible voting decisions s/he makes on the other bills? In order to answer this question, please sort the y column in the Distribution Table in descending order. Please choose your answers in HW2 Submission.

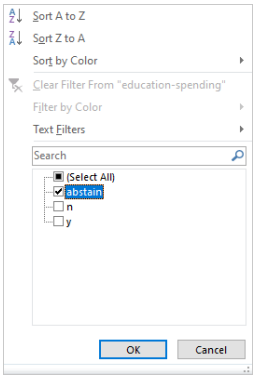
5 Modeling: Logistic Regression

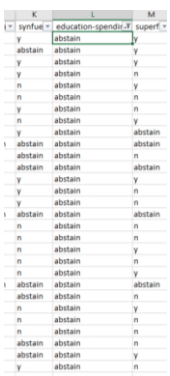
5.1 Take a look at our data, you will observe that our target attribute is a polynominal attribute because it has more than two categories, y, n, and abstain. Therefore, the first issue of logistic regression (DV has more than two categories) is involved in this case. Because you are asked to predict whether the "education-spending" Bill introduced to the Congress will pass or not, you are interested in the category y. Accordingly, in this lab, we use the solution 2 mentioned on PPT Slide 28. How to put this solution into practice? There are at least two methods (Method 2 is easier).

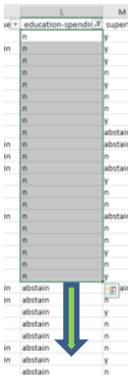
Method 1: Change abstain to n in Excel and then reimport the dataset to RM (Step 5.2).

5.2 Data Preparation in Excel and Import it to RM in a new process

5.2.1 Open the training dataset in your Excel and then apply Filter to Column L, education-spending. Select “abstain” and then change all the “abstain” to “n” (Attention: for some software package, you may have to change your categories to 1-0 or Yes-No). Alternatively, you can do this in RM, but it takes a few more steps (please find the process in the appendix).





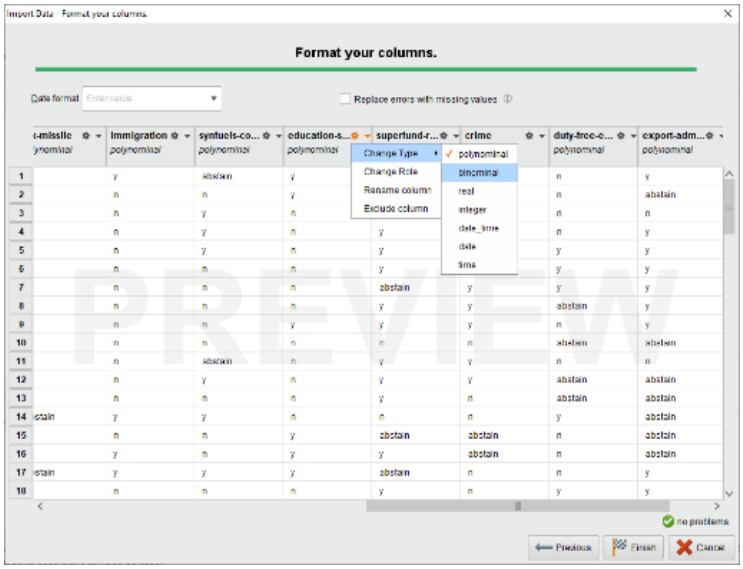


5.2.2 Save your dataset as vote-train – LR.csv.

5.2.3 Import the new dataset via the Read CSV operator. At the third step of the Import Configuration Wizard, Format your columns, please change data type of educational-spending from polynominal to binominal because we have only two categories now.

5.2.4 Set educational-spending as the target attribute as we did before.

5.2.5 Add a new operator Logistic Regression. Note: please do not use the other two logistic regression operators such as Logistic Regression (SVM).

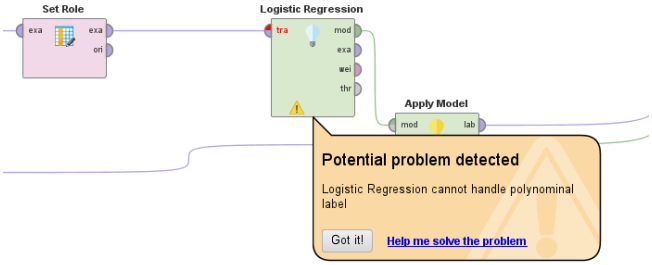


Method 2: Replace abstain to n in RM directly (Step 5.3).

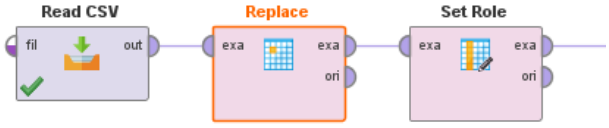
5.3 Work on the process at Step 3.2.

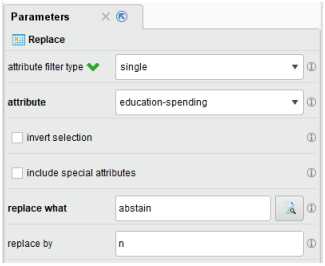
5.3.1 Replace Naïve Bayes by the operator Logistic Regression

5.3.2 Logistic Regression operator gives you an error message: Logistic Regression cannot handle polynominal label.



5.3.3 Add a new operator "Replace" between Read CSV and Set Role, and set its parameters as below.





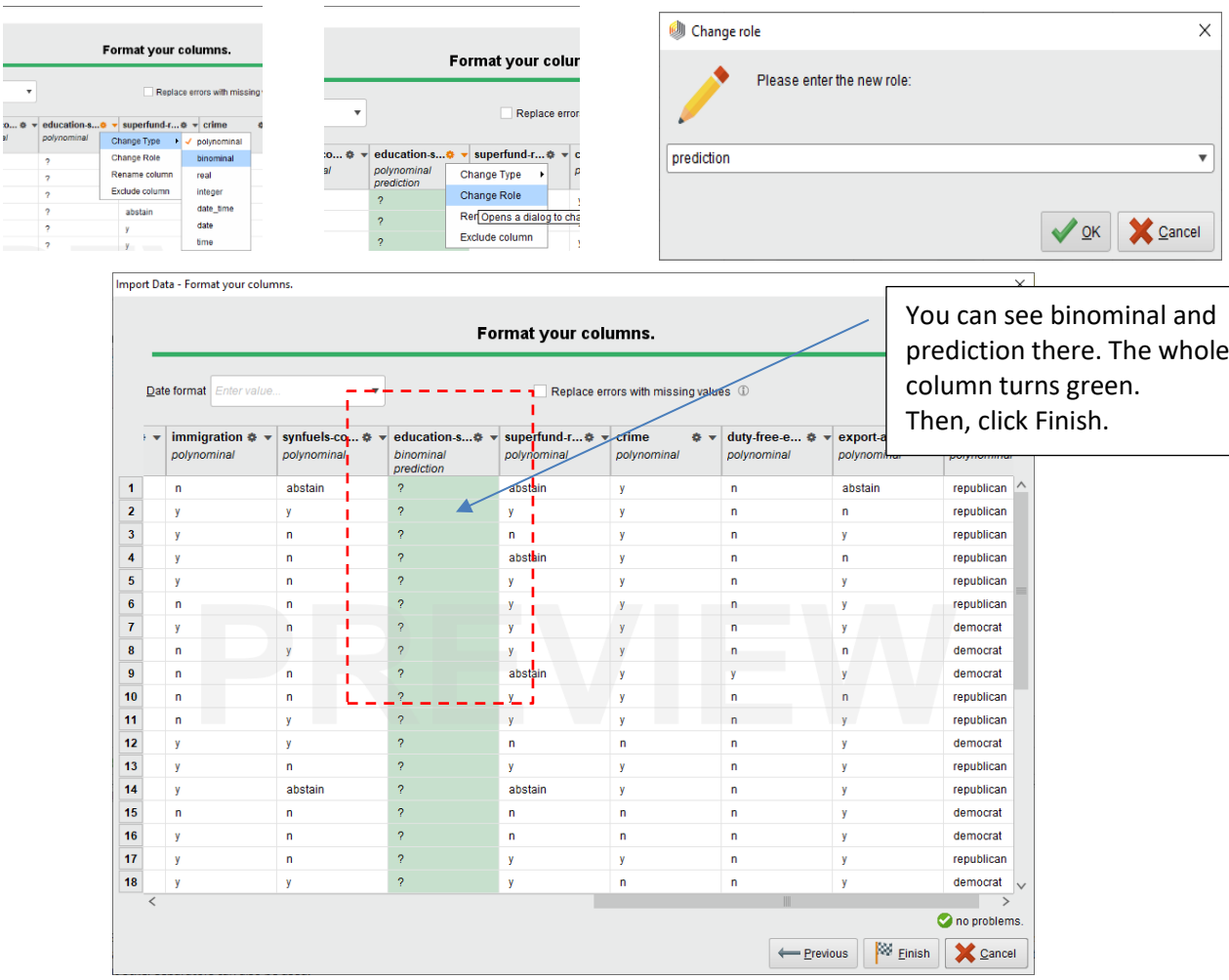
5.3.4 Even the error message of Logistic Regression operator still exists, you can ignore it because the target attribute now has only two values, even though its variable type has not been changed to binominal. If you want to change it to binomial, please refer to the process in the appendix.

5.4 You can either use method 1 or 2. Save your process as Week5_LR and then run it. Take a look at the Statistic view of your prediction results and then answer the questions: How many people are predicted to vote "y" (i.e., Yes)? Type a whole number here. Would the Bill pass the Congress? Type Yes or No here.

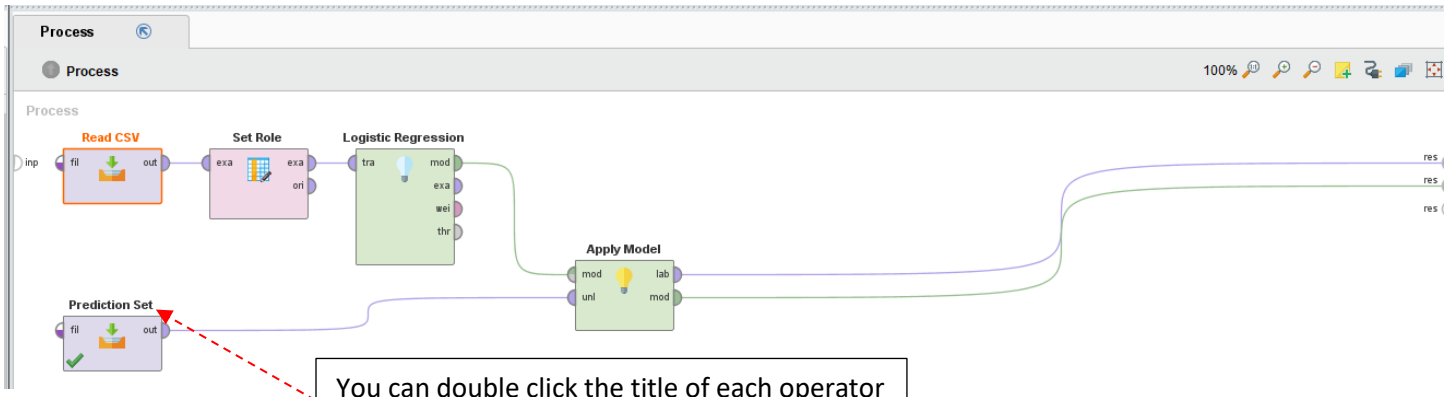
Another tip is described as below to help you set target attribute differently.

When importing the prediction dataset via the Read CSV operator. At the third step of the Import Configuration Wizard, Format your columns, please make the following two changes:

- Change Type of educational-spending from polynominal to binominal (Optional).
- Change Role to prediction. Doing so, we do not need to use the operator Set Role for the prediction. After you make this change, you will see that this column turns green.



You can also double click the prediction set and change its name from Read CSV to Prediction or Prediction Set.



6 Logistic Regression Interpretation

6.1 When you view the output of logistic regression model, you find that RM automatically convert categorical predictor attributes to dummy attributes. For example, the original attribute crime is converted to two dummy attributes: crime.abstain and crime.n (see the table in the right).

Original Attribute	Coded to New Dummy Attributes	
crime	crime.abstain	crime.n
abstain	1	0
n	0	1
y	0	0

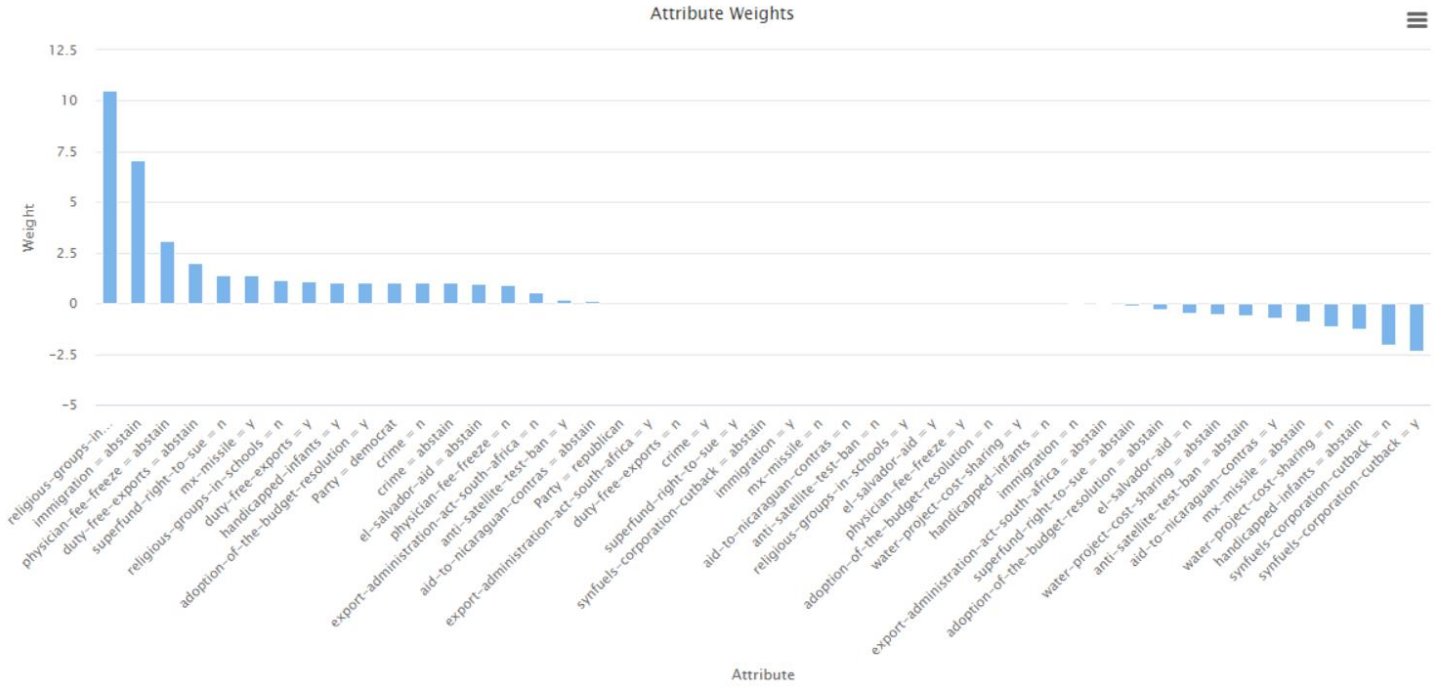
6.2 You can also take a look at those regression coefficient.

Result History | LogisticRegression (Logistic Regression) | ExampleSet (Apply Model)

Data | Description | Annotations

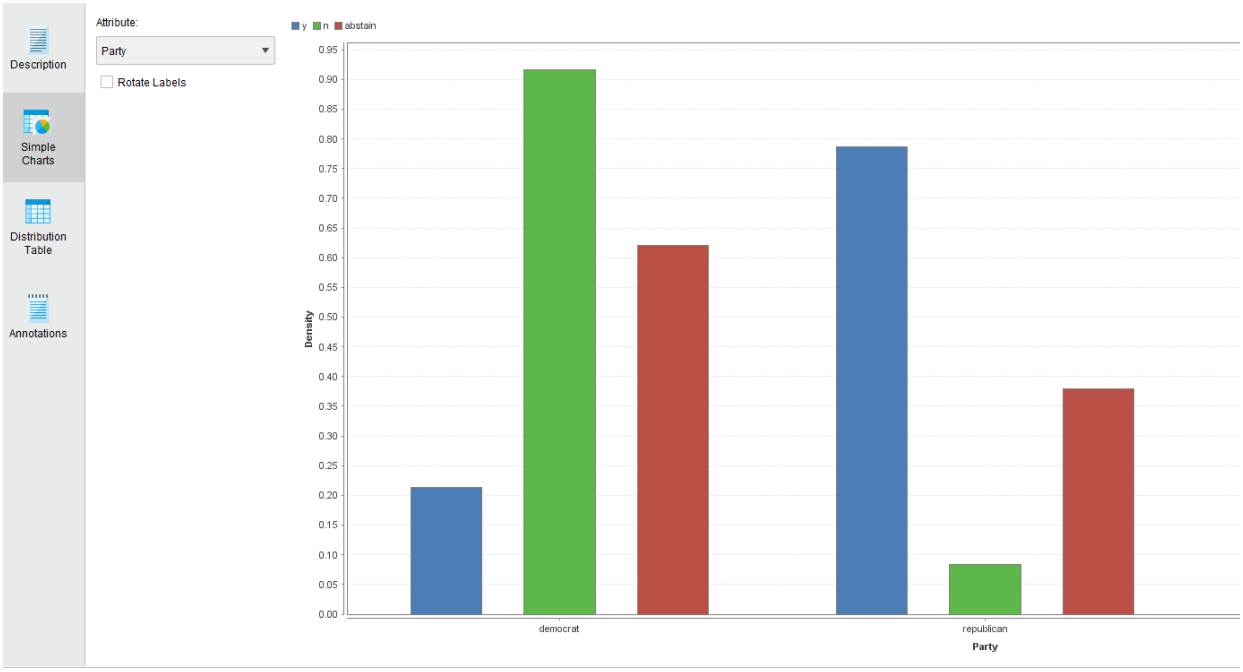
Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
Intercept	-0.549	-0.549	1.054	-0.521	0.602
Party.democrat	1.014	1.014	0.750	1.351	0.177
adoption-of-the-budget-resolution.abstain	-0.300	-0.300	1.309	-0.229	0.819
adoption-of-the-budget-resolution.y	1.017	1.017	0.474	2.147	0.032
aid-to-nicaraguan-contras.abstain	0.133	0.133	1.166	0.114	0.909
aid-to-nicaraguan-contras.y	-0.695	-0.695	0.650	-1.070	0.284
anti-satellite-test-ban.abstain	-0.602	-0.602	1.074	-0.560	0.575
anti-satellite-test-ban.y	0.175	0.175	0.500	0.351	0.726
crime.abstain	1.010	1.010	1.271	0.795	0.427
crime.n	1.011	1.011	0.580	1.745	0.081
duty-free-exports.abstain	1.973	1.973	0.806	2.448	0.014
duty-free-exports.y	1.096	1.096	0.451	2.429	0.015
el-salvador-aid.abstain	0.950	0.950	1.297	0.733	0.464
el-salvador-aid.n	-0.503	-0.503	0.735	-0.684	0.494
export-administration-act-south-africa.abstain	-0.079	-0.079	0.480	-0.164	0.870
export-administration-act-south-africa.n	0.544	0.544	0.535	1.017	0.309
handicapped-infants.abstain	-1.266	-1.266	1.655	-0.765	0.444
handicapped-infants.y	1.028	1.028	0.420	2.449	0.014
immigration.abstain	7.086	7.086	148.063	0.048	0.962
immigration.n	-0.035	-0.035	0.375	-0.095	0.925
mx-missile.abstain	-0.876	-0.876	0.812	-1.079	0.281

6.3 You can choose to generate another result by drawing a line between the wei port of Logistic Regression and res port. Then, you can generate an Attribute Weights column chart as below.



6.4 Please note that instead of running the logistic regression for the positive class ("y"), RM runs it for the

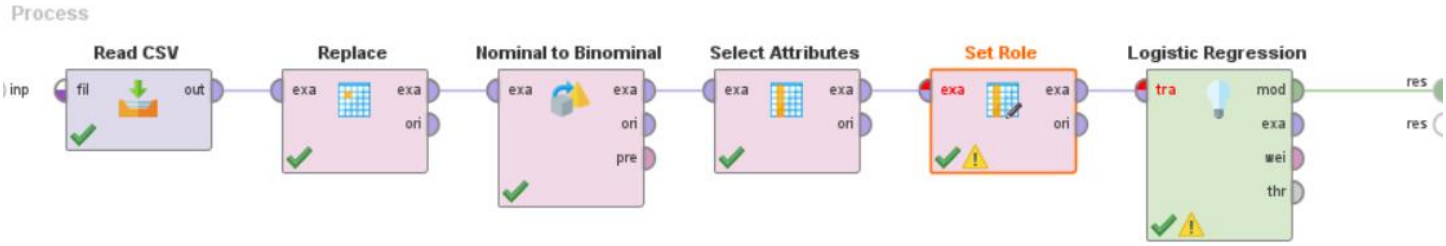
negative class (R runs the logistic regression for the positive class). Accordingly, when interpreting the regression coefficients (or weights), we need to take the opposite direction of those regression coefficients. For example, the regression coefficient of Party.democrat is positive, which indicates that it positively predicts the probability of voting n (no) and negatively predicts the probability of voting y (yes) in the bill of education-spending. If a Congressman converts from a Democrat (1) to Republican (0), the log odds of voting y increases by 1.014. This further indicates that the odds of voting yes is $e^{1.104}$ =3.02 times higher for a Republican than for a Democrat. This is quite consistent with what we get from Naïve Bayes. Republicans are more likely to vote yes on education-spending than Democrats.



Conditional Probabilities Generated by the Naïve Bayes Model

6.5 Take a look at the regression coefficients and answer the question: which attributes listed in in HW2 Submission have a negative and significant relationship with the probability of voting "y" in the bill of education-spending? Note: we use p<0.05 as the criterion to determine that a regression coefficient in significant. Hint: you can sort those regression coefficients by p value in an ascending order.

Appendix: Replacing education-spending=abstain and Changing its datatype to Binominal in RM



Replace abstain by n	Change polynominal to binominal	Unselect education-spending = n	Set education-spending = y as target attribute	Run logistic regression

If you use this process to generate a logistic regression model, all the regression coefficients (or weights) represent the correct relationship with the positive case (i.e., “y” in this case) because we require RM to use education-spending = y as target attribute as the target attribute.