**Lab: How to prepare data for the training set**

Download the excel sheet for the data set named as “raw-customer-churn-data” from model. Follow the steps to prepare the data for the training set.

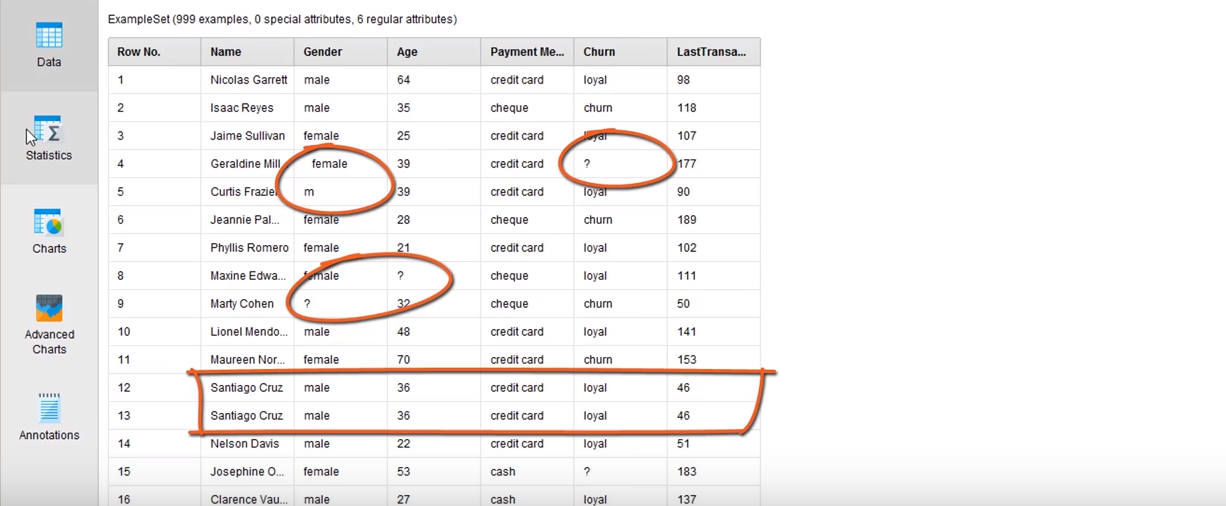
Preparation of Data from Raw Data

1) Drag the Read excel operator in the process window.

2) Open the file "raw-customer-churn-data" from the parameter panel.

3) Link the input or output ports.

4) Check the Statistics of data and you can observer two major things as highlighted in the figure.



5) Missing values for the attributes called as Churn (Recall last week exercise) and Age have missing values, whereas the attribute Gender has four set of values **(female, male, m, ?)**.

6) Two major categories for Gender (male and female), but we found four because of missing data and errors in the data. The predictive analytics can be performed based on data sets that contains significant errors.

7) Analytics tasks must be completed before doing any modelling.

8) We use “**trim operator”** that applied to all attributes and It is responsible to remove the white spaces in the data set.

9) Then we use "**Filter Example**" operator to perform further operations

10) Now we apply filter on various attributes

*11) Use "Import Configuration" visit from the parameter window to get all attributes on Add filter condition.*

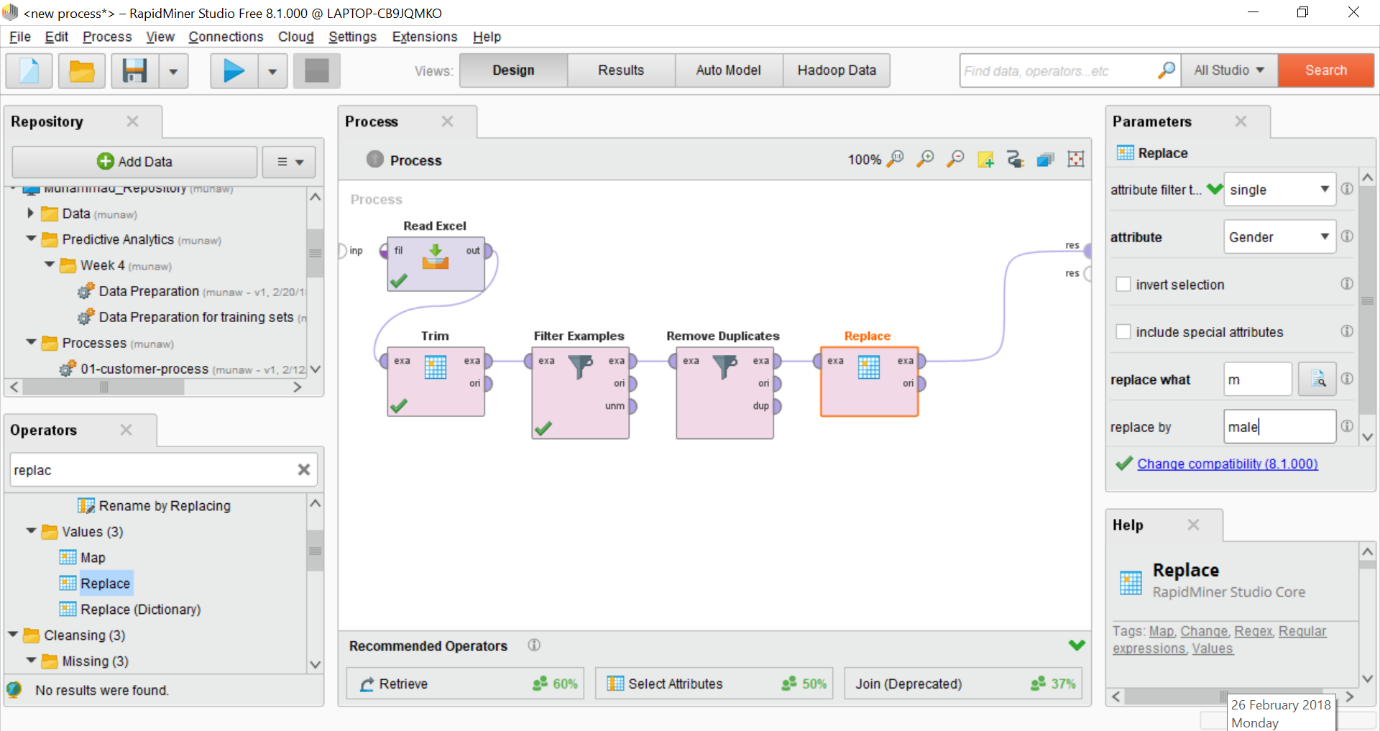
12) For Gender attribute, select the filter "is not missing"

13) For Age attribute, select the filter "is not missing"

14) Next Operator is “**removal duplicate”** on the current process window and select all attributes.

15) If you would like to replace "m" with "male", we need a “**replace operator”** to perform this action.

16) Change some values on the parameter pane to implement the action mentioned on the line 15 and you can get help from the screen shot as mentioned below



17) we also use the **regular expression** to complete this action, replace the letter "m" with "\b[m]\b", "\b" means beginning and end of the word.

18) The further details of regular expression are available on this website, "www.regular-expression.info"

19) Now we need to split the data into two columns for the predictive analytics.

20) Apply the “**filter example”** operator again to perform these actions

21) Select the churn attribute as "is not missing" by using filter operator.

22) Now we use “**store operator”** to store all data set values at our desired location.

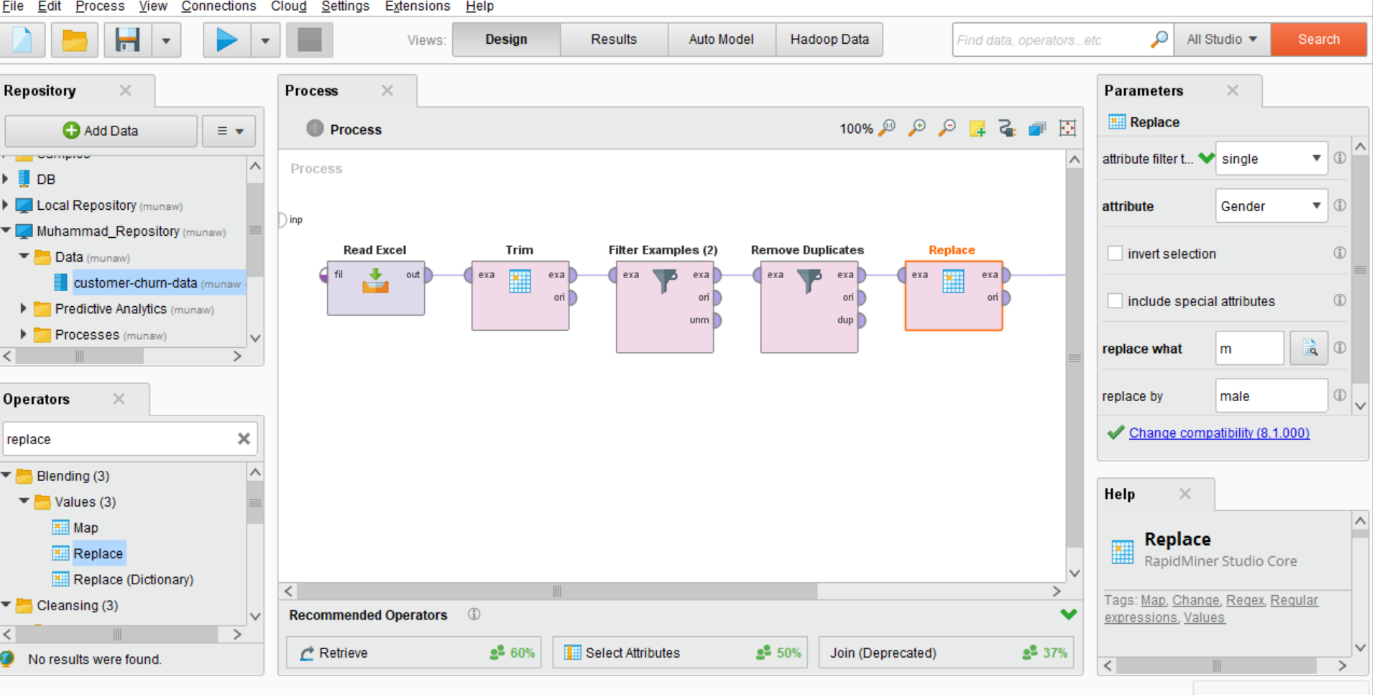
23) Link the operator **Filter example** port "unum" to the **store operator** port "inp", then select the location for the storage of data file. Store in the hard drive repository and named as "Unlabeled customers".

24) After the second Filter example operator, apply the operator "**Select attribute**" and use a single label for the churn attribute. Also tick the invert selection option.

25) As there are number of operators present in the process window, we do some housekeeping.

26) We can move all cleaning processes into sub window

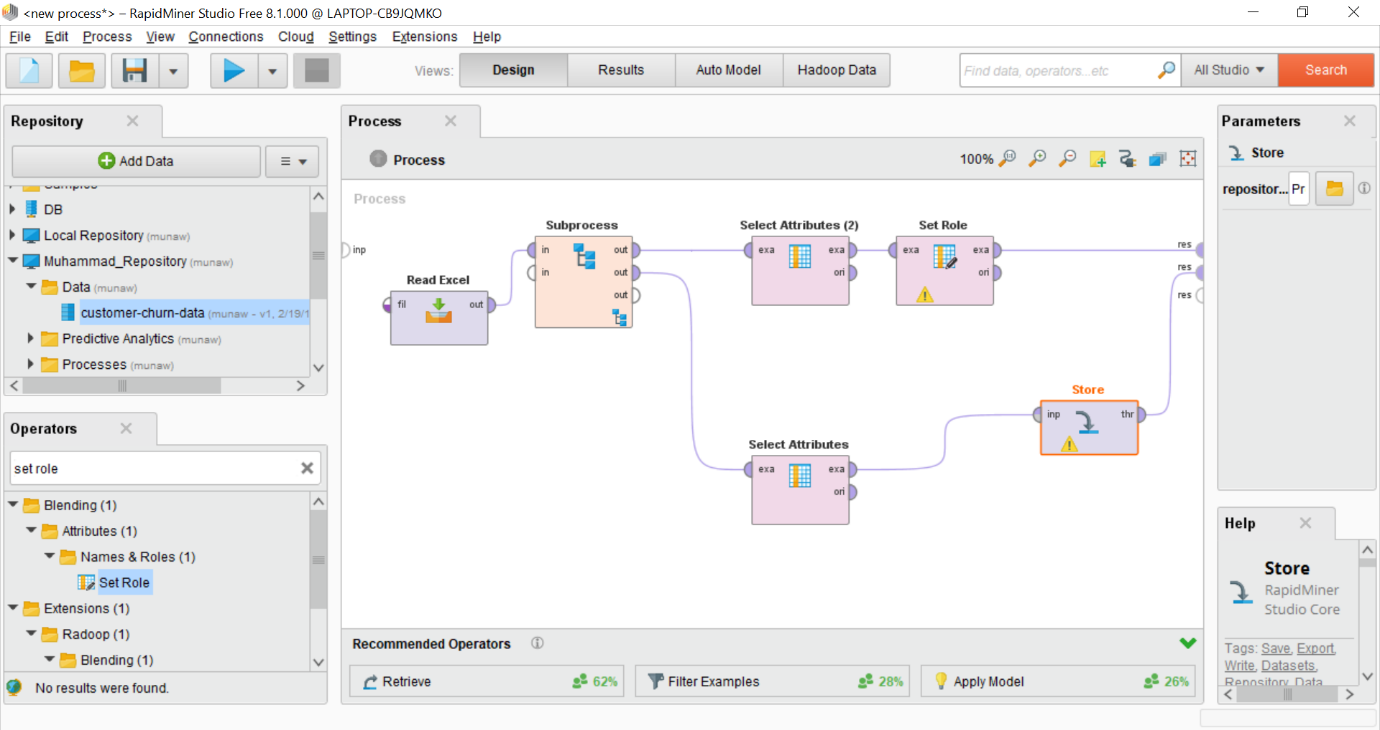
27) Select all the operators in the process window, then right click and select on the option for subprocess and we can give a descriptive name if you like.



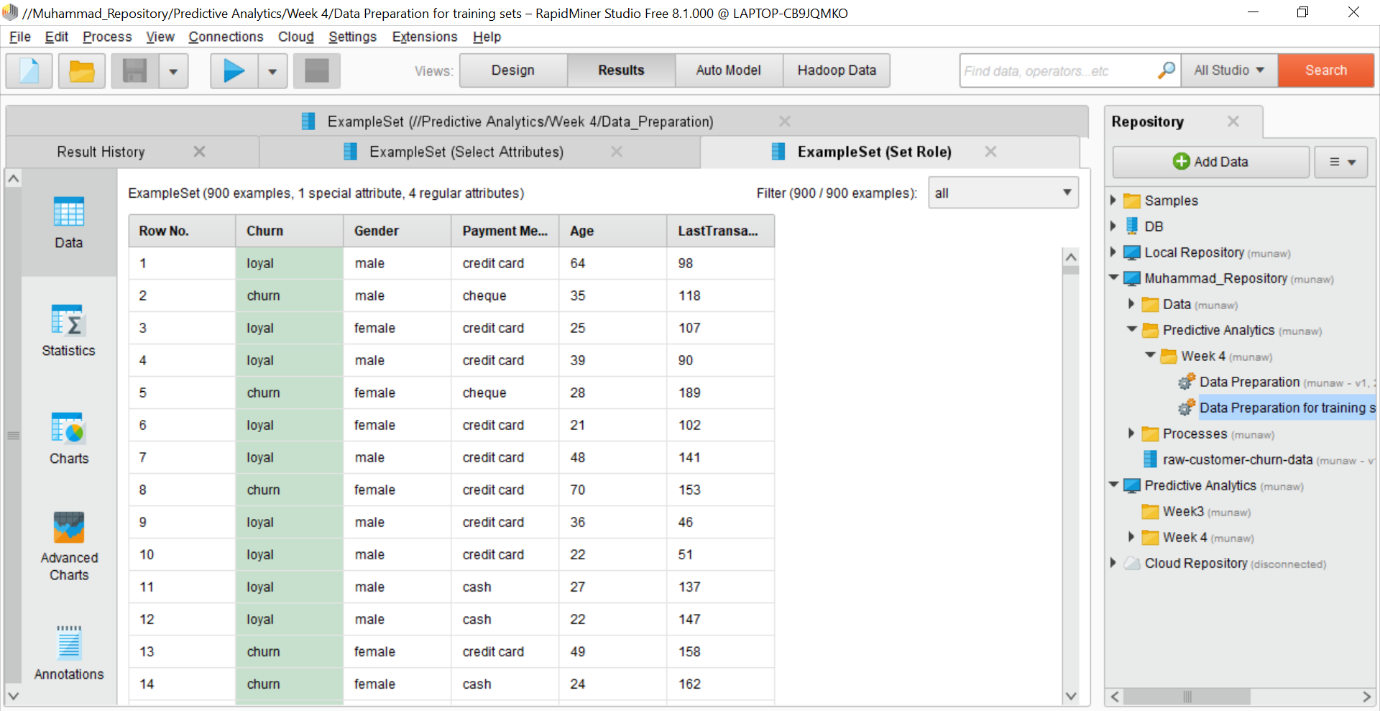
28) Before the final development of training data

29) We select the "Select attribute" variable before the final target. We consider "Name" as single attribute and tick for invert selection.

30) For Churn attribute, we use "**Set role**" operator and target role select as "label".

31) Finally, we connect the store operator to the result port before the final run.

32) You can see the two data sets for the training and you can now apply model on that training data set.



* If you like to explore further, you can watch a youtube video as mentioned below
* <https://www.youtube.com/watch?v=vzV3OT2h1Gg&list=PLssWC2d9JhOZLbQNZ80uOxLypglgWqbJA&index=6>