RapidMiner Project Example

Telco Customer Churn Prediction (data from Kaggle)

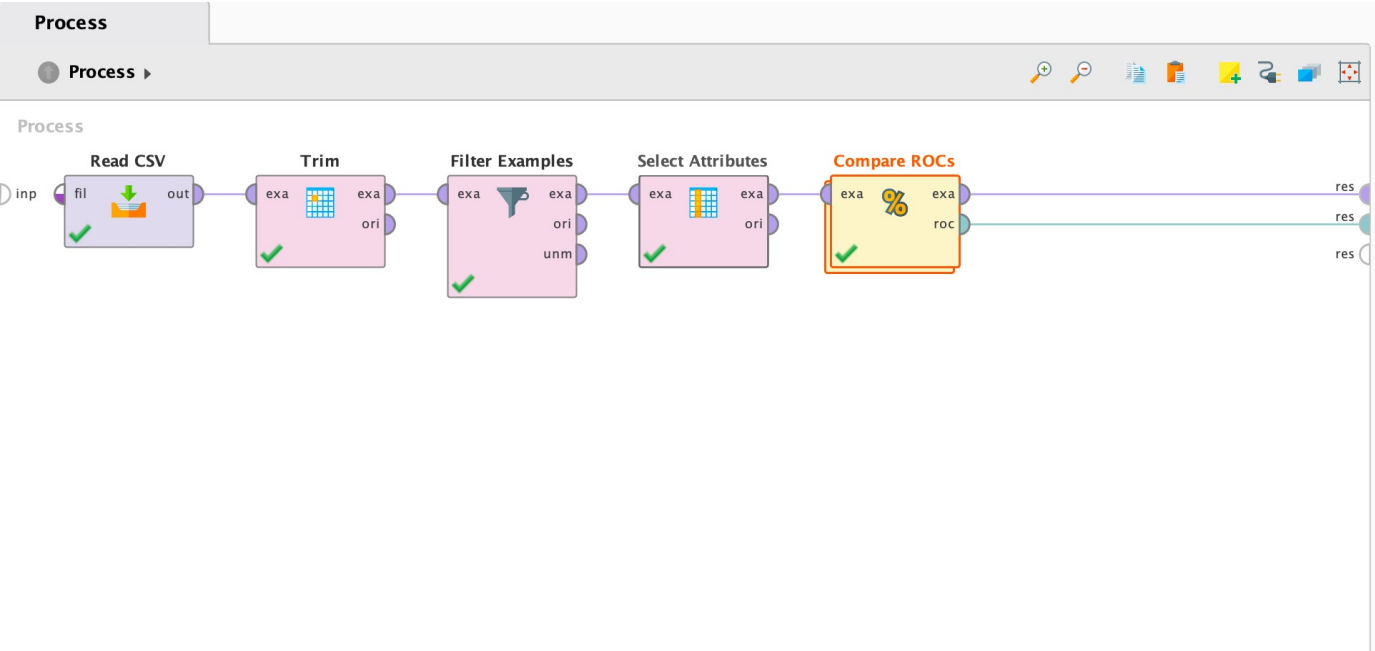
Create by Michael Kim

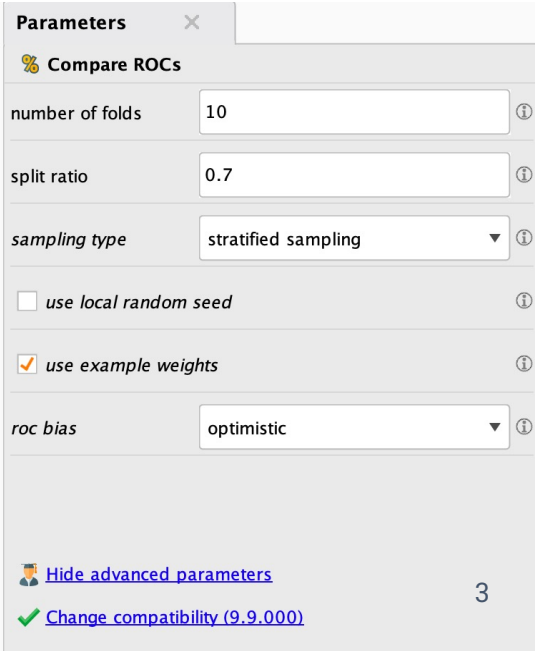
Data Preparation

- The data for this assignment has been downloaded from Kaggle (https://www.kaggle.com/blastchar/telco-customer-churn).
- I have split the data into a training dataset and a test dataset (7:3 ratio).
- Churn (Yes/No) is the ***label*** column.

Data Preparation & Cleaning

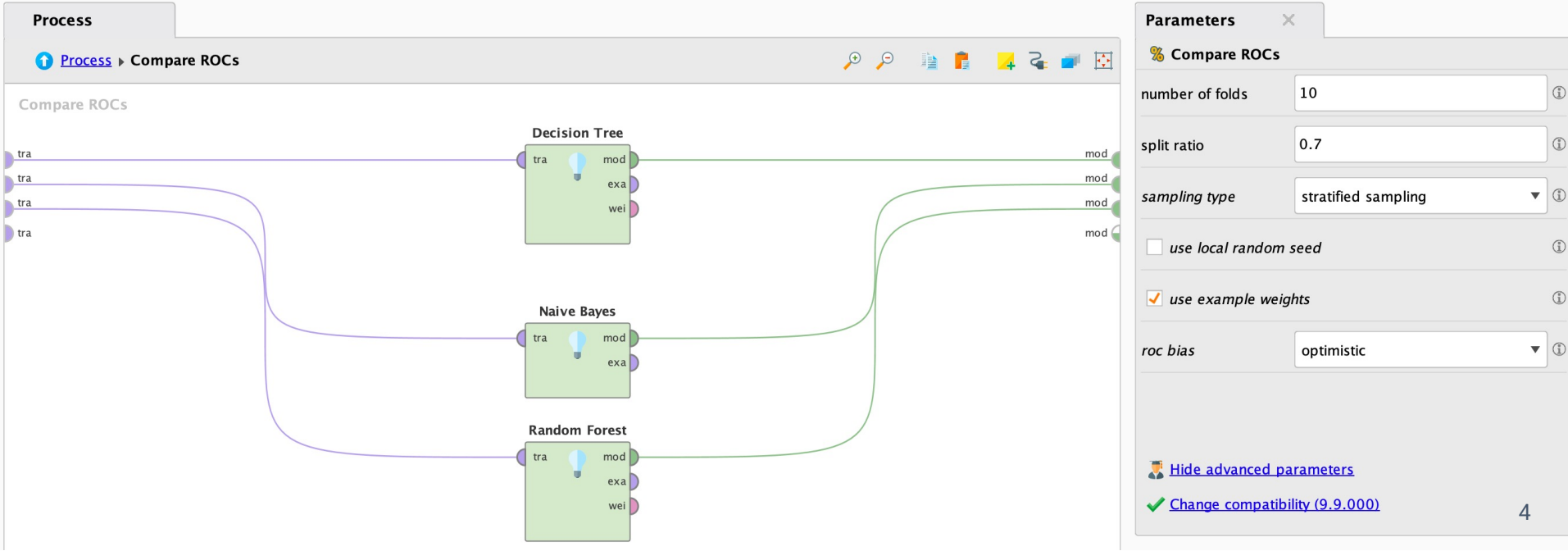
- The data is very clean, except that there are a few missing values in TotalCharges column.
- Used '*Trim*' & '*Filter Examples*' operator to get rid of missing values, and '*Select Attributes*' operator to filter out customerID column.





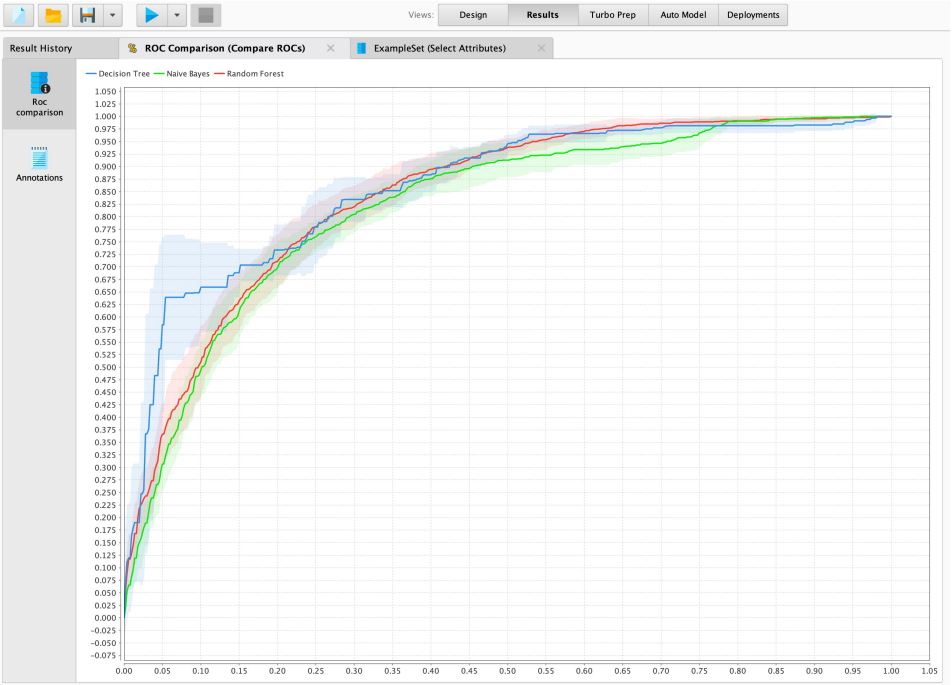
Model Selection: Classification

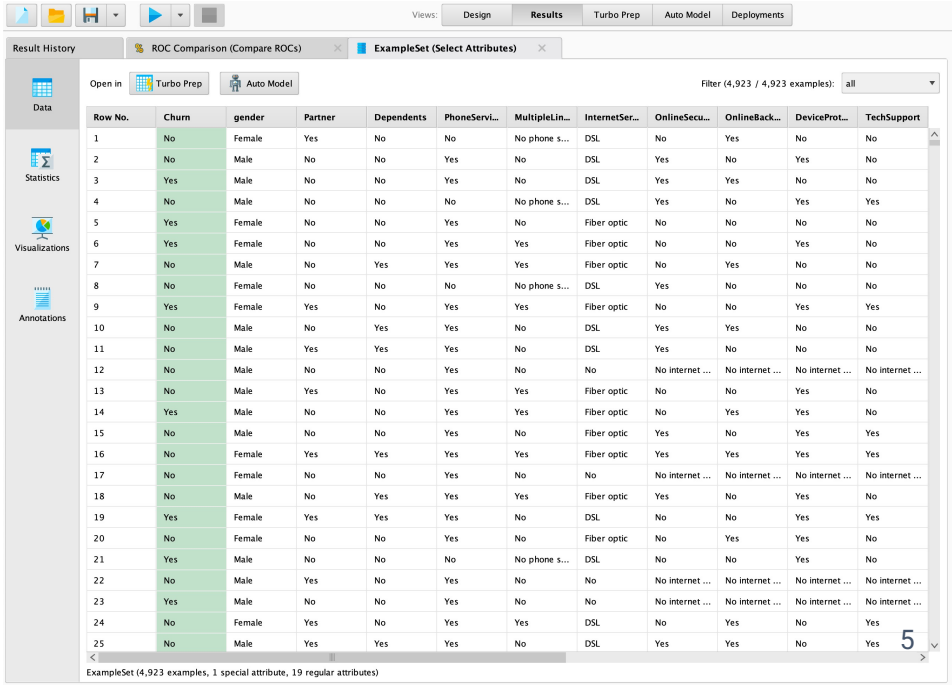
- Used '*Compare ROCs'* operator to determine the best classification algorithms.
- Within 'Compare ROCs', I put *Decision Tree*, *Random Forest* and *Naïve Bayes*.



ROC Curve

- The closer an ROC curve is to the upper left corner, the more efficient is the model.
- ***Decision Tree*** is the best model developed for this data.

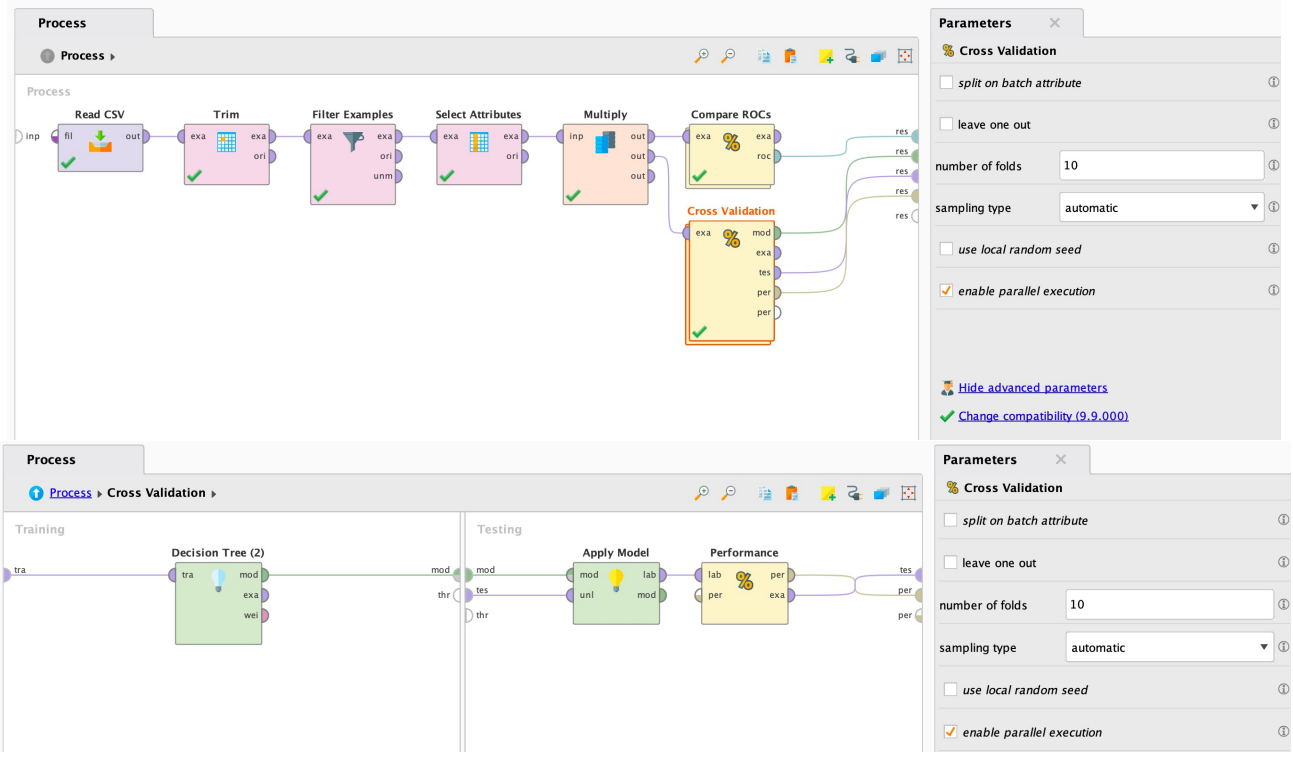




Row No.	Churn	gender	Partner	Dependents	PhoneServi...	MultipleLin...	InternetSer...	OnlineSecu...	OnlineBack...	DeviceProt...	TechSupport
1	No	Female	Yes	No	No	No phone s...	DSL	No	Yes	No	No
2	No	Male	No	No	Yes	No	DSL	Yes	No	Yes	No
3	Yes	Male	No	No	Yes	No	DSL	Yes	Yes	No	No
4	No	Male	No	No	No	No phone s...	DSL	Yes	No	Yes	Yes
5	Yes	Female	No	No	Yes	No	Fiber optic	No	No	No	No
6	Yes	Female	No	No	Yes	Yes	Fiber optic	No	No	Yes	No
7	No	Male	No	Yes	Yes	Yes	Fiber optic	No	Yes	No	No
8	No	Female	No	No	No	No phone s...	DSL	Yes	No	No	No
9	Yes	Female	Yes	No	Yes	Yes	Fiber optic	No	No	Yes	Yes
10	No	Male	No	Yes	Yes	No	DSL	Yes	Yes	No	No
11	No	Male	Yes	Yes	Yes	No	DSL	Yes	No	No	No
12	No	Male	No	No	Yes	No	No	No internet ...	No internet ...	No internet ...	No internet ...
13	No	Male	Yes	No	Yes	Yes	Fiber optic	No	No	Yes	No
14	Yes	Male	No	No	Yes	Yes	Fiber optic	No	Yes	Yes	No
15	No	Male	No	No	Yes	No	Fiber optic	Yes	No	Yes	Yes
16	No	Female	Yes	Yes	Yes	Yes	Fiber optic	Yes	Yes	Yes	Yes
17	No	Female	No	No	Yes	No	No	No internet ...	No internet ...	No internet ...	No internet ...
18	No	Male	No	Yes	Yes	Yes	Fiber optic	Yes	No	Yes	No
19	Yes	Female	Yes	Yes	Yes	No	DSL	No	No	Yes	Yes
20	No	Female	No	No	Yes	No	Fiber optic	No	Yes	Yes	No
21	Yes	Male	No	No	No	No phone s...	DSL	No	No	Yes	No
22	No	Male	Yes	No	Yes	No	No	No internet ...	No internet ...	No internet ...	No internet ...
23	Yes	Male	No	No	Yes	No	No	No internet ...	No internet ...	No internet ...	No internet ...
24	No	Female	Yes	No	Yes	Yes	DSL	No	Yes	No	Yes
25	No	Male	Yes	Yes	Yes	No	DSL	Yes	Yes	No	Yes

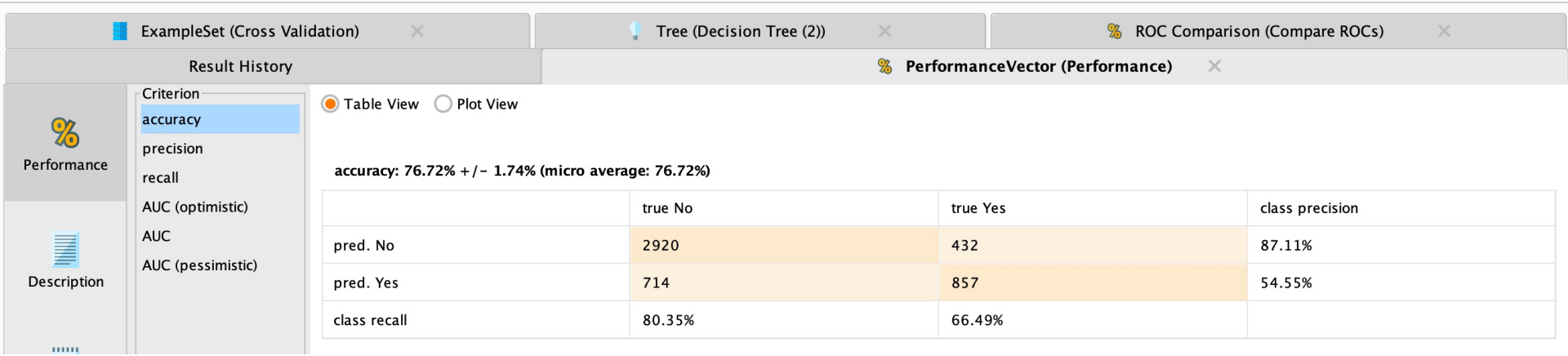
Model Evaluation with Cross Validation

- To measure the model performance, ‘Cross Validation’ operator is used. In this process, *Decision Tree* algorithm is used.



Decision Tree Performance

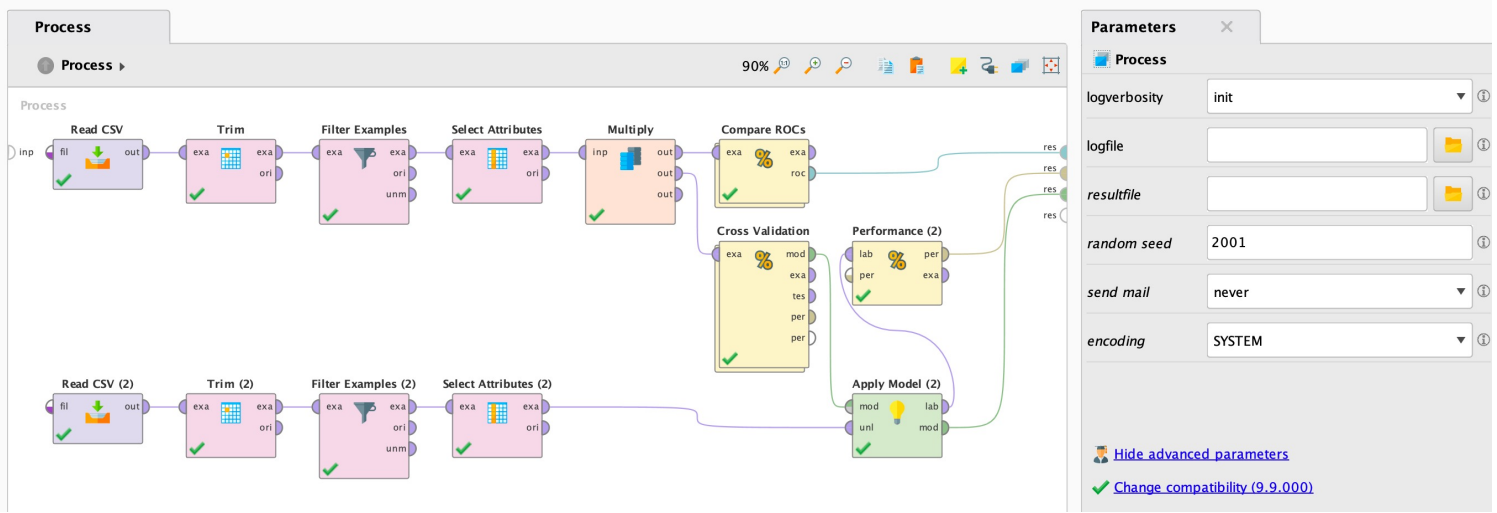
- The accuracy of the model: **76.72%**



	true No	true Yes	class precision
pred. No	2920	432	87.11%
pred. Yes	714	857	54.55%
class recall	80.35%	66.49%	

Model Evaluation

- Applied the model on the *test* dataset, and the accuracy is **75.34%**



accuracy: 75.34%

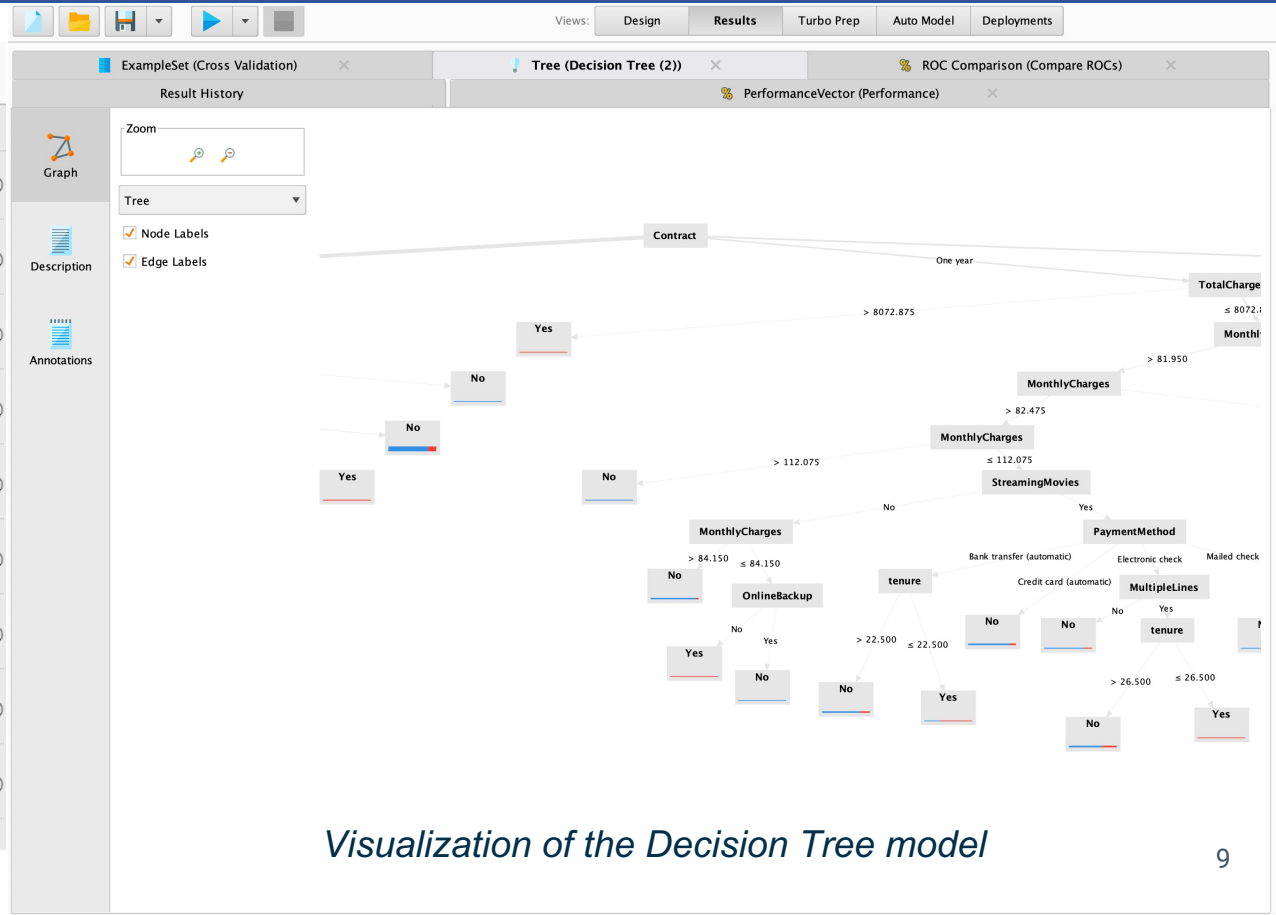
	true No	true Yes	class precision
pred. No	1212	203	85.65%
pred. Yes	317	377	54.32%
class recall	79.27%	65.00%	

Decision Tree: Overview & Visualization

Parameters

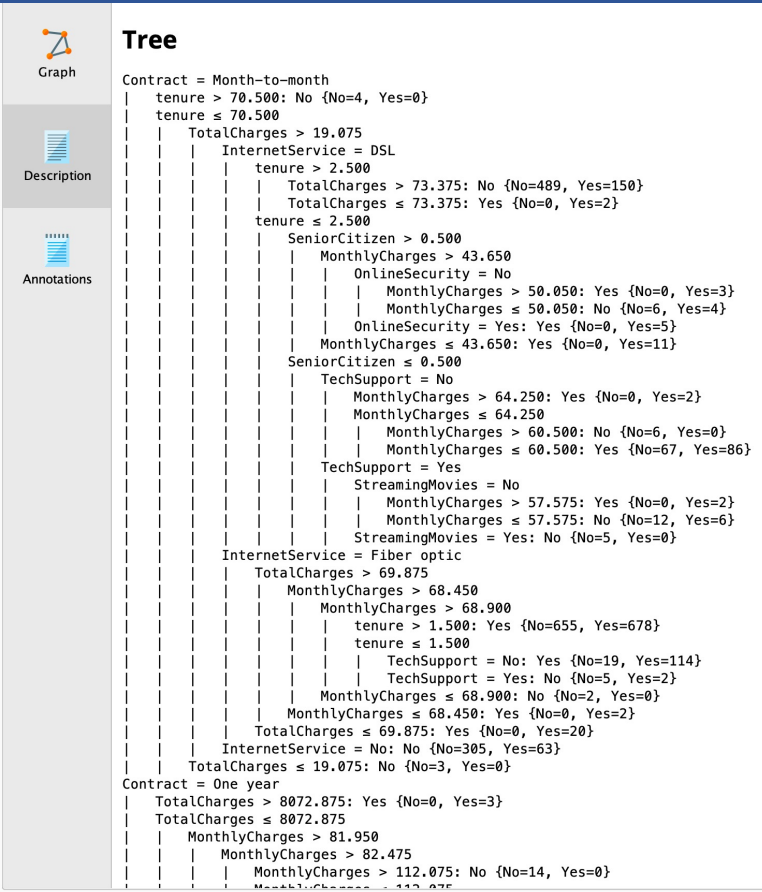
Decision Tree (2) (Decision Tree)

criterion	gain_ratio
maximal depth	10
apply pruning	☑
confidence	0.1
apply prepruning	☑
minimal gain	0.01
minimal leaf size	2
minimal size for split	4
number of prepruning alternatives	3



Visualization of the Decision Tree model

Decision Tree: Description & Important Features



According to the description of the Decision Tree model, the most important factors that determine the customer churn are:

- Contract
- TotalCharges
- Tenure
- MonthlyCharges
- InternetService
- SeniorCitizen

Conclusion

- The purpose of the data modeling is to predict the customer churn.
- *Decision Tree*, *Random Forest* and *Naïve Bayes has been used to determine the best classification model, and* ***Decision Tree*** is the best model developed for this data.
- To measure the model performance, 'Cross Validation' operator is used and *Decision Tree* algorithm is used in this process. The accuracy of the model is **76.72%.**
- When the model is applied on the *test* dataset, and the accuracy is **75.34%**
- The most important factors that determine the customer churn are: *Contract*, *TotalCharges*, *Tenure*, *MonthlyCharges*, *InternetService*, and *SeniorCitizen.*

RapidMiner Project Example

Thank you!