PROJECT REPORT ON DATA MINING TO DETECT EMPLOYEE ATTRITION

Using RapidMiner

Prepared by: TEAM 6

Brad Thele

Keka Nandi

Samtha Reddy

Virali Pathak

Department of Information Systems
NEW JERSEY INSTITUTE OF TECHNOLOGY, NEWARK

ABSTRACT

Attrition is a big issue for companies today. Attrition is the reduction of employees in a business through regular means, such as resignation and retirement. Many employees currently believe that staying with one company for a long time isn't worth it since they could change jobs more frequently and make more money over time. This has made it important for companies today to keep track of their attrition rate to make sure that their rate of turnover is normal, and that they aren't losing too many employees to continue running their business effectively.

Our project used a data set that gave us metrics about employees at a company. Our goal was to use data mining methodologies to determine if there is a correlation between some of these metrics and attrition. By using data mining we will be able to determine what makes an employee more or less likely to leave a company.

Table of Contents

1 INTRODUCTION

In today's competitive business environment, the impact of attrition on a business can be detrimental to both the bottom line and morale. To decrease attrition, managers must understand the causes of customer and employee turnover, the costs associated with attrition, and finally, institute measures to reduce attrition rates.

Employee Attrition

Causes of employee attrition can be as varied as human personalities, but some basic factors pervade most reasons for a resignation. While employees leave an employer for increased salary or career advancement, often the search for a new position is precipitated by dissatisfaction with an immediate manager. An employee's personality might not be a good fit for the job. New skills for an employee's current position could be needed. If the employee lacks the skill to do the job, the employee might resign.

Costs of Employee Attrition

The costs of employee attrition range from quantifiable numbers to hidden costs. When employees resign from companies, costs are incurred in recruiting new employees and training them. Productivity will be lower until new hires learn the business. If it is a customer-based business, customers could be dissatisfied if the new hire is not proficient. The business could lose customers who are dissatisfied with service. Revenue would decrease.

Solution Modelling

Classification is the data mining technique that involves the use of supervised machine learning techniques, which predict the categorical class labels. In addition, its accuracy depends on the percentage of the test samples that are correctly classified.

In this study, we focus on a fictional dataset created by IBM data scientist which was hosted in Kaggle.com and apply two data mining algorithms to predict the most accurate model to detect employees who might leave or stay with the company.

Data Preparation

The dataset contained 35 attributes out of which we cleansed and used only 23 attributes. The attributes deleted were: DailyRate, DistanceFromHome, EmployeeCount, EmployeeNumber, Over18, MonthlyRate, Standard Hours, etc. These attributes mostly had same values for the entire dataset or the values were irrelevant.

The resultant dataset of 23 attributes had 1470 records and did not contain any null or junk values for the attributes. Also no data transformation was required.

The attributes considered were: Age, Attrition, BusinessTravel, Department, Education, EducationField, EnvironmentSatisfaction, Gender, JobInvolvement, JobLevel, JobRole, JobSatisfaction, MaritalStatus, NumCompaniesWorked, OverTime, PercentSalaryHike, PerformanceRating, TotalWorkingYears, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager. **Attrition** is our class variable and has 2 values viz. yes=that the employee has left and no=that the employee is retained.

Methodology

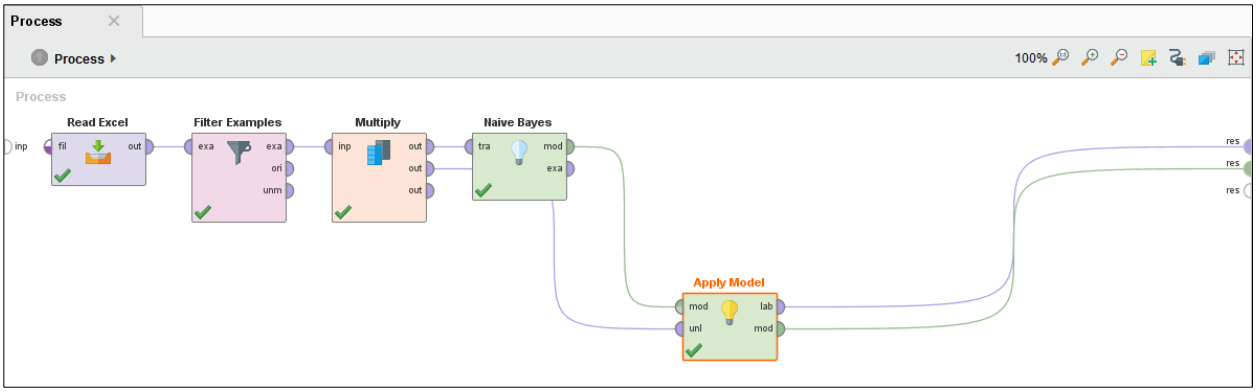
Data Mining is the process of extracting information from a huge set of data. According to the Wikipedia, the overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Similarly, here in this study also we try to formulate predictions based on available dataset i.e. focus on predictive data mining.

In order to achieve the learning value from this dataset we use Naïve Bayes and Decision Tree as the Data Mining Technique and our aim is to compare the results and obtain the most suitable solution. To achieve the same, we have used the Rapid Miner tool, which is a software platform that provides an integrated environment for data mining, predictive analytics and is used for business, commercial applications as well as for research, education and training.

2 Naïve Bayes

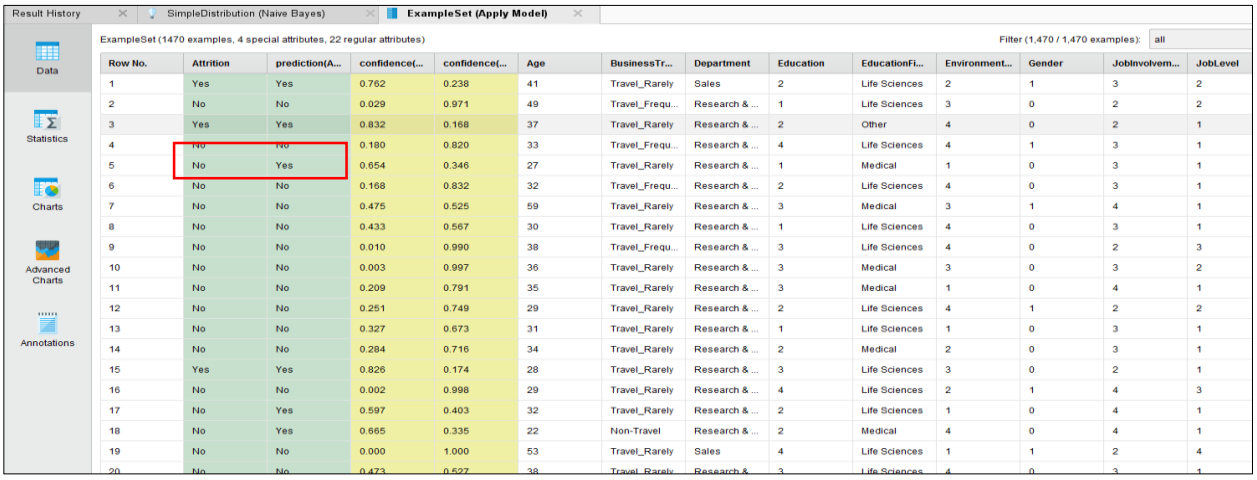
Naïve Bayes is a very useful approach because it yields strong results quickly. It assumes that each piece is unrelated to the other pieces (which is not true - but it works anyways) and then finds correlation. Naïve Bayes is great for large data sets and although it is simple it has proven itself as a very reliable method over the years.

2.1 RapidMiner Model



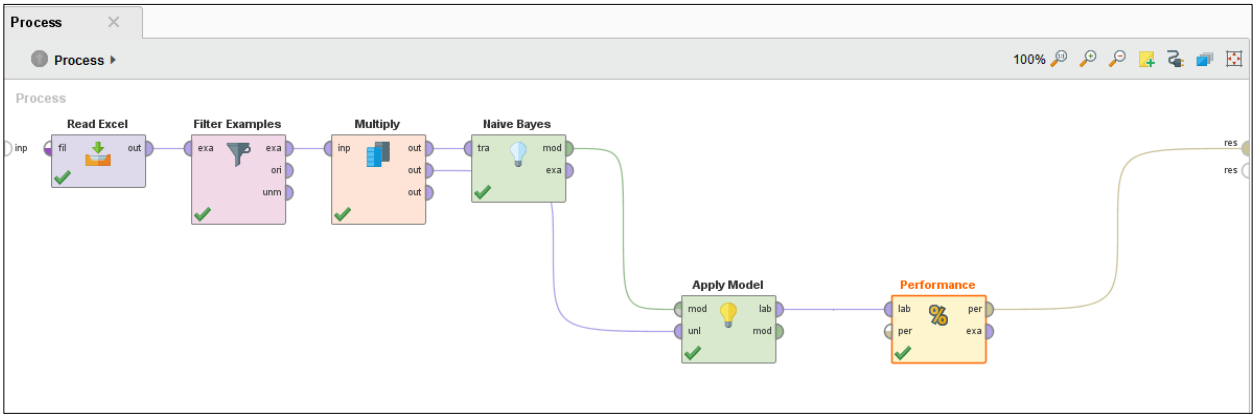
Rapid Miner Results:

We observe, some of the predications were incorrect.

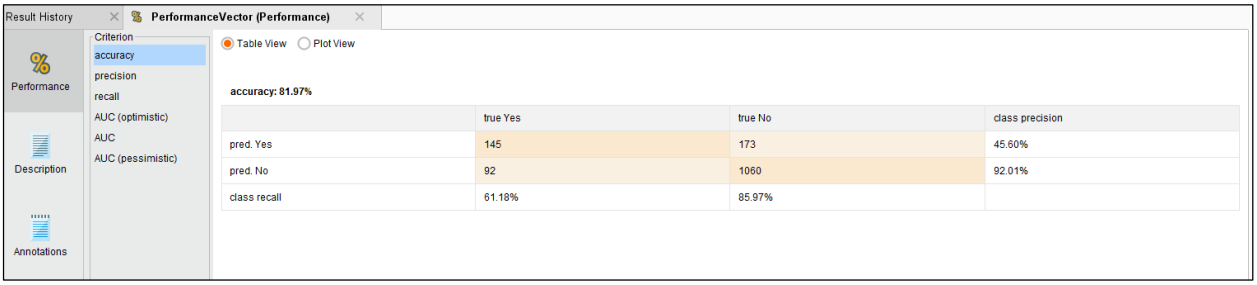


Row No.	Attrition	prediction(A...	confidence(...	confidence(...	Age	BusinessTr...	Department	Education	EducationFi...	Environment...	Gender	JobInvolvem...	JobLevel
1	Yes	Yes	0.762	0.238	41	Travel_Rarely	Sales	2	Life Sciences	2	1	3	2
2	No	No	0.029	0.971	49	Travel_Frequ...	Research & ...	1	Life Sciences	3	0	2	2
3	Yes	Yes	0.832	0.168	37	Travel_Rarely	Research & ...	2	Other	4	0	2	1
4	No	No	0.180	0.820	33	Travel_Frequ...	Research & ...	4	Life Sciences	4	1	3	1
5	No	Yes	0.654	0.346	27	Travel_Rarely	Research & ...	1	Medical	1	0	3	1
6	No	No	0.168	0.832	32	Travel_Frequ...	Research & ...	2	Life Sciences	4	0	3	1
7	No	No	0.475	0.525	59	Travel_Rarely	Research & ...	3	Medical	3	1	4	1
8	No	No	0.433	0.567	30	Travel_Rarely	Research & ...	1	Life Sciences	4	0	3	1
9	No	No	0.010	0.990	38	Travel_Frequ...	Research & ...	3	Life Sciences	4	0	2	3
10	No	No	0.003	0.997	36	Travel_Rarely	Research & ...	3	Medical	3	0	3	2
11	No	No	0.209	0.791	35	Travel_Rarely	Research & ...	3	Medical	1	0	4	1
12	No	No	0.251	0.749	29	Travel_Rarely	Research & ...	2	Life Sciences	4	1	2	2
13	No	No	0.327	0.673	31	Travel_Rarely	Research & ...	1	Life Sciences	1	0	3	1
14	No	No	0.284	0.716	34	Travel_Rarely	Research & ...	2	Medical	2	0	3	1
15	Yes	Yes	0.826	0.174	28	Travel_Rarely	Research & ...	3	Life Sciences	3	0	2	1
16	No	No	0.002	0.998	29	Travel_Rarely	Research & ...	4	Life Sciences	2	1	4	3
17	No	Yes	0.597	0.403	32	Travel_Rarely	Research & ...	2	Life Sciences	1	0	4	1
18	No	Yes	0.665	0.335	22	Non-Travel	Research & ...	2	Medical	4	0	4	1
19	No	No	0.000	1.000	53	Travel_Rarely	Sales	4	Life Sciences	1	1	2	4
20	No	No	0.473	0.527	38	Travel_Rarely	Research & ...	3	Life Sciences	4	0	3	1

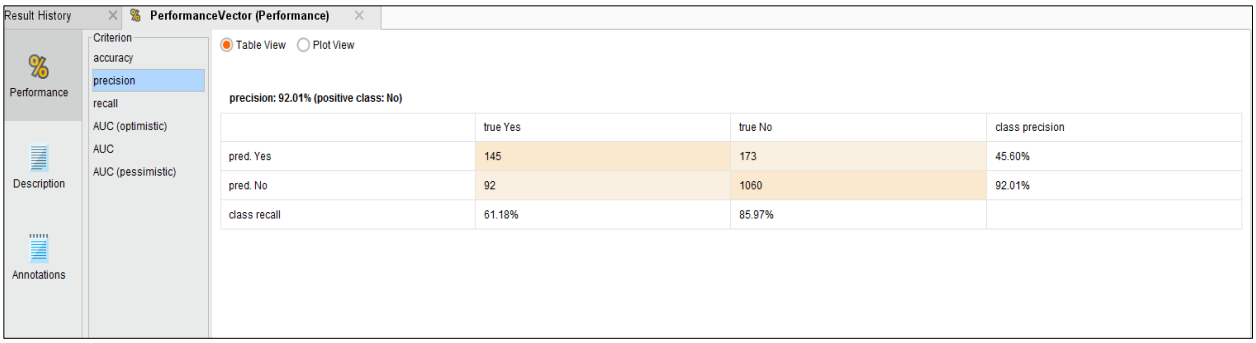
Applying perfomance operator to measure accuracy of “Naïve Bayes” model for the given dataset.

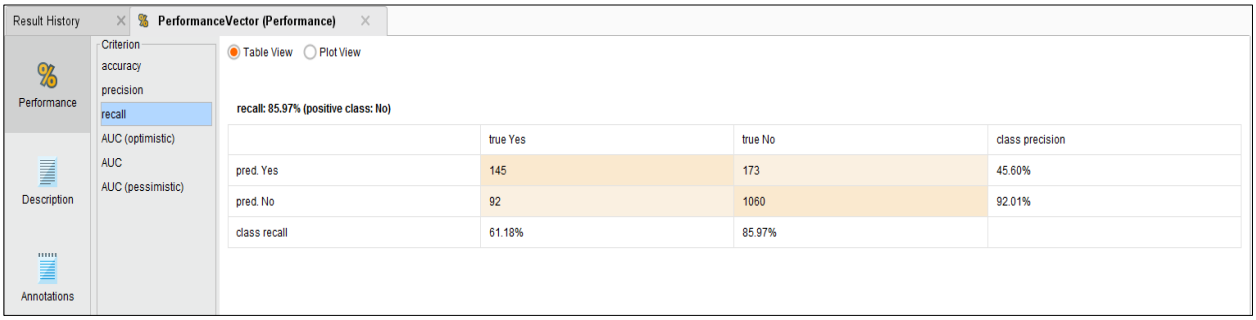


2.2 Confusion Matrix

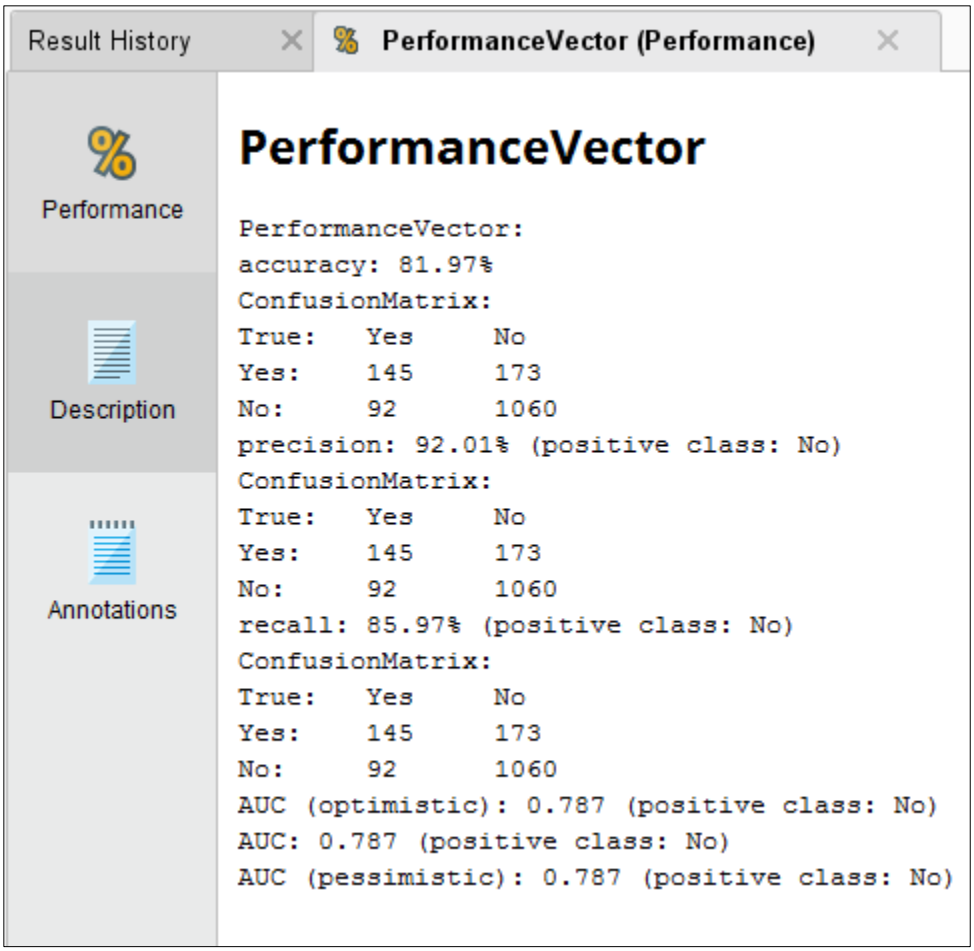


	true Yes	true No	class precision
pred. Yes	145	173	45.60%
pred. No	92	1060	92.01%
class recall	61.18%	85.97%	





Summary of Confusion Matrix: Note: Here positive class: No.

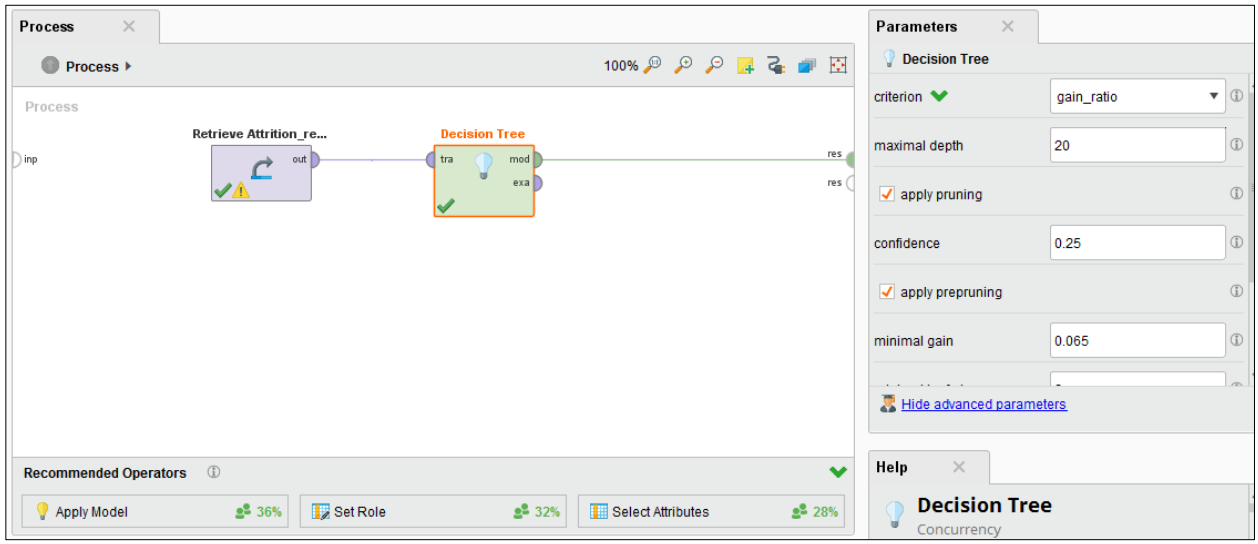


3 Decision Tree

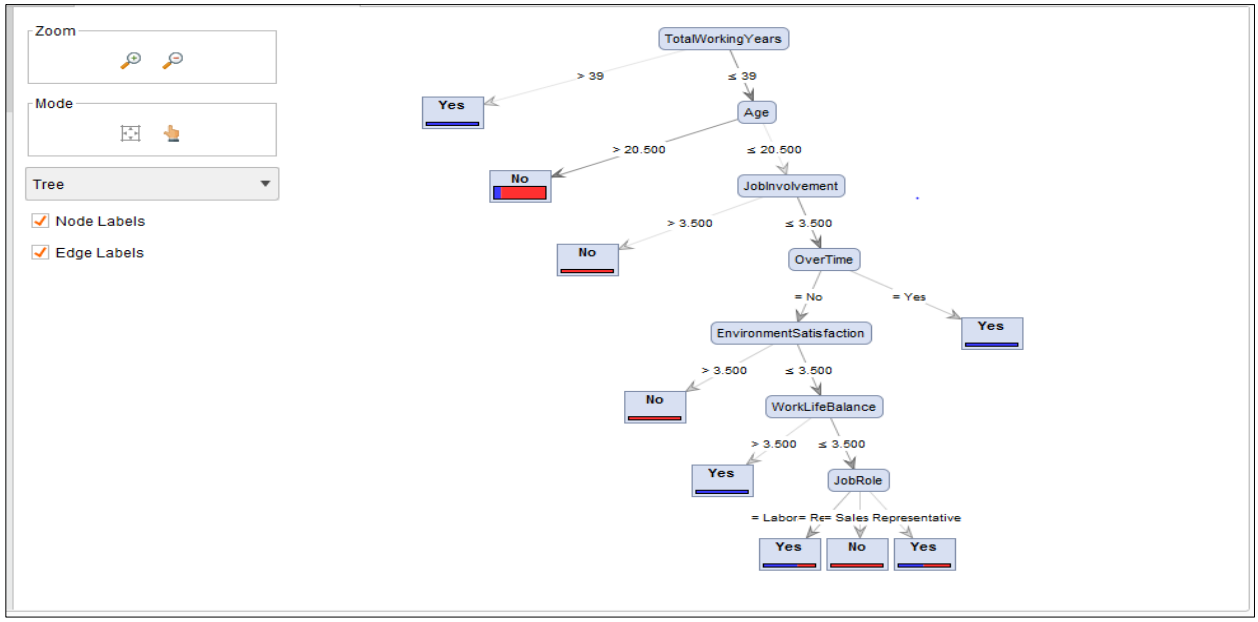
Decision tree model is a simple but powerful form of multiple variable analysis. This multiple variable analysis capability of decision trees enables us to go beyond one cause, one effect relationships and discover and describe things in context of multiple influences. It is the most popular methodology because of its ease-of-use, robustness with a variety of data and levels of measurement and ease of interpretability.

3.1 RapidMiner Model

The dataset was retrieved and Decision Tree operator was applied in RapidMiner. The minimal gain was changed from the default value of 0.1 to 0.065 to get an optimal tree.

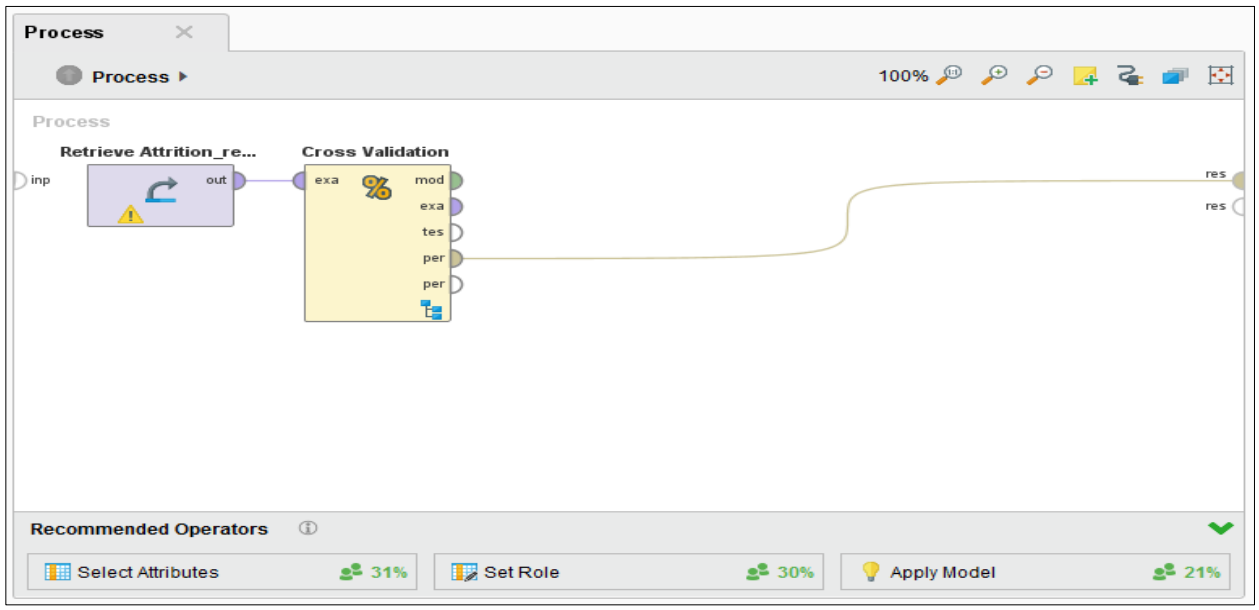


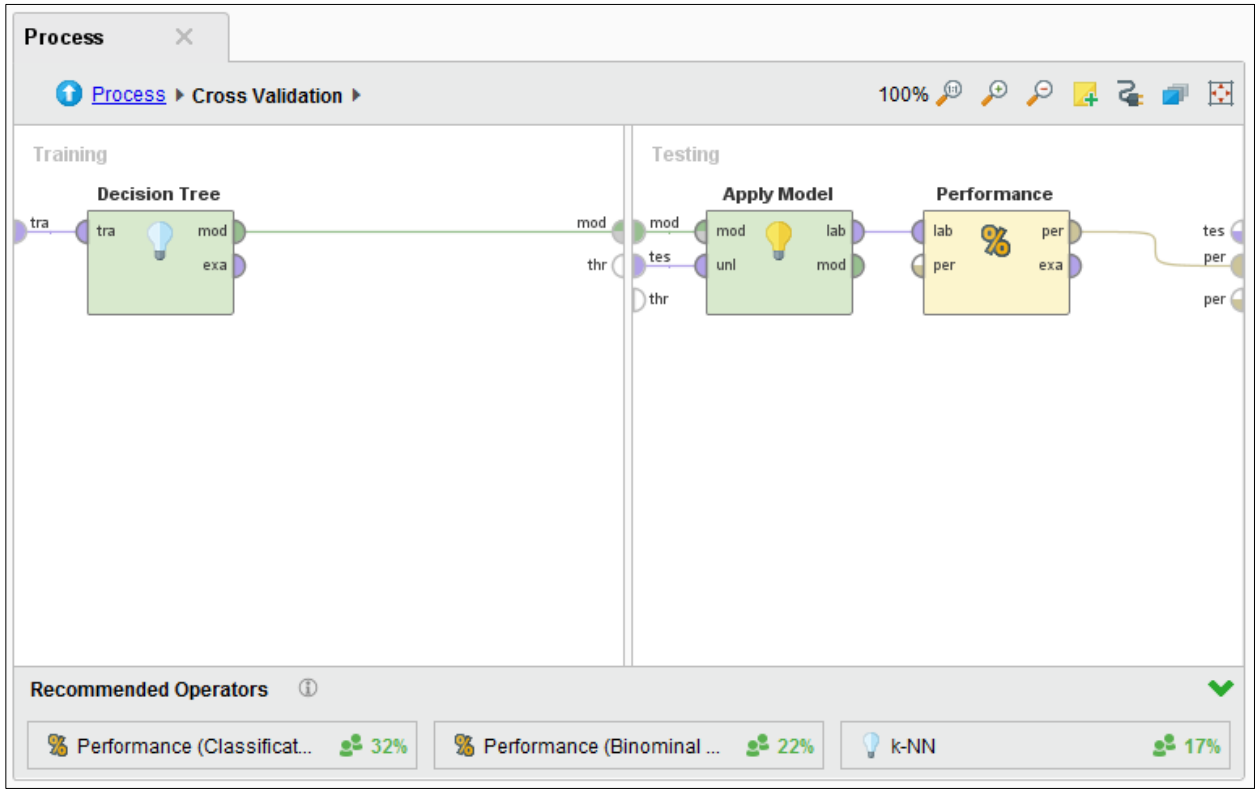
Decision Tree obtained is:



3.2 Cross Validation of Model

The cross validation operator is applied to the dataset with the default number of folds = 10.





3.3 Confusion Matrix

The results obtained from cross validation are as below:

accuracy: 83.95% +/- 0.62% (mikro: 83.95%)

	true Yes	true No	class precision
pred. Yes	24	23	51.06%
pred. No	213	1210	85.03%
class recall	10.13%	98.13%	

precision: 85.04% +/- 0.69% (mikro: 85.03%) (positive class: No)

	true Yes	true No	class precision
pred. Yes	24	23	51.06%
pred. No	213	1210	85.03%
class recall	10.13%	98.13%	

recall: 98.14% +/- 1.36% (mikro: 98.13%) (positive class: No)

	true Yes	true No	class precision
pred. Yes	24	23	51.06%
pred. No	213	1210	85.03%
class recall	10.13%	98.13%	

PerformanceVector

```
PerformanceVector:
accuracy: 83.95% +/- 0.62% (mikro: 83.95%)
ConfusionMatrix:
True:    Yes     No
Yes:     24      23
No:      213     1210
precision: 85.04% +/- 0.69% (mikro: 85.03%) (positive class: No)
ConfusionMatrix:
True:    Yes     No
Yes:     24      23
No:      213     1210
recall: 98.14% +/- 1.36% (mikro: 98.13%) (positive class: No)
ConfusionMatrix:
True:    Yes     No
Yes:     24      23
No:      213     1210
AUC (optimistic): 0.932 +/- 0.036 (mikro: 0.932) (positive class: No)
AUC: 0.525 +/- 0.032 (mikro: 0.525) (positive class: No)
AUC (pessimistic): 0.119 +/- 0.054 (mikro: 0.119) (positive class: No)
```

4 Result Comparison

After applying two mining techniques on the same dataset, the results are summarized as below:

(Note: The positive class is NO i.e. the employee who has not left)

Mining Method	Accuracy	Precision	Recall
Naïve Bayes	81.97%	92.01%	85.97%
Decision Tree	83.97%	85.04%	98.14%

As per the classification models, below are the few questions answered:

a) How often is our classifiers correct ?
Approximately 82-84% for both the models Naïve Bayes & Decision Tree i.e in both cases, out of 1470 records, the classifier had classified (True positives + True negatives) = 1205 and 1234 times respectively.
So, as per accuracy, Decision tree is a better classifier.

b) Out of those employees who, in fact,stayed (label: No) or left (label: Yes), what proportion was classified that way by the model? (Recall)
For Yes scenarios (i.e. employees who left), the Naïve Bayes predicted Yes = 61.18% and for No = 85.97%
For Yes scenarios, the decision tree predicted Yes = 10.13% and for No = 98.14%
The recall is better in Naïve Bayes considering both the % of Yes & No whereas decision tree is better if only % of No is considered.

c) Out of those employees who were predicted to stay (label: No) or leave (label: Yes), what proportion actually stayed or left ? (Precision)
For Yes scenarios (i.e. employees who left), the Naïve Bayes precision was = 45.60% and for No scenarios = 92.01%
For Yes scenarios, the decision tree precision was = 51.06% and for No = 85.03%
The precision is better in Decision Tree.

5 Conclusion

Our aim was to create a model that predicts whether or not an employee will attrite based on our known variables. There are two possible predicted classes: "yes" and "no".

The essence of confusion matrix for attrition scenario can be summarized as below, taking class “No” as positive class.

True Positive(TP): Employees predicted as no and they in reality did not leave the company.
False Positive(FP): Employees predicted as no and they in reality did leave the company.
True Negative(TN): Employees predicted as Yes and they in reality did leave the company.
False Negative(FN) : Employees predicted as Yes and they in reality did not leave the company.

	No(True)	**Yes(True)**
No(Pred)	TP	FP (Type I error)
Yes(Pred)	FN(Type II error)	TN

Recall	TP/(TP+FN)
Precision	TP/(TP+FP)

An organization, would like to know, out of all the employees, who would continue to work in their company for longer number of years. HR would be more interested in hiring an candidate whose likelyhood of staying is high, than whose chances of attrition is less. Thus we want a model which can get more **true positive cases.**

Also a company would be concerned if the perentage of attrition in current employees is higher. It makes more sense for a company to detect all employees who would actually leave, even if in that process we incorrectly predicting some employees as “Yes”, though they didn’t leave. Thus we want a model which can **high value of TN, so that chances of Type II error are less**.

Hence we need model which has better recall at the cost of precision and the recall for Naïve Bayes is better for both cases where employees stay (label: No) and leave (label: Yes) rather than Decision tree model where the recall for only employees who stay back (label: No) is good.

6 Future Scope

When an organization can correctly identify the employees who tend to be loyal, they can reward them with annual gift vouchers and credit points. Also the organization can take feedback from these valued employees to improve the retention rate and convert the employees to stay back who might have been predicted to leave. As mentioned earlier, attrition in a company incurs a lot of cost and brand defamation. Rather, the organization can spend the same or lesser cost in retaining an employee by rewarding them and increase the retention rate.

7 Limitation & Assumptions

The modelling was done on a fictitious dataset created by IBM data scientist and the labels: "Yes" or "No" were assumed to mean employees who leave and stay respectively. We did not have any metadata description of the label. Also the proportion of "No" records was higher in the dataset, which is why the models in the RapidMiner took the positive models to be: No and displayed the precision & recall value.

8 References

- http://support.sas.com/publishing/pubcat/chaps/57587.pdf
- Tutorials at RapidMiner, Inc. (Getting Started with the RapidMiner Platform)
- Lecture Notes from IS 665
- https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset
- http://www.marketwatch.com/story/i-worked-here-20-years-and-all-i-got-was-a-blender-2013-09-12
- http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/
- http://docs.statwing.com/the-confusion-matrix-and-the-precision-recall-tradeoff/