Data mining using

CRISP-DM
CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING

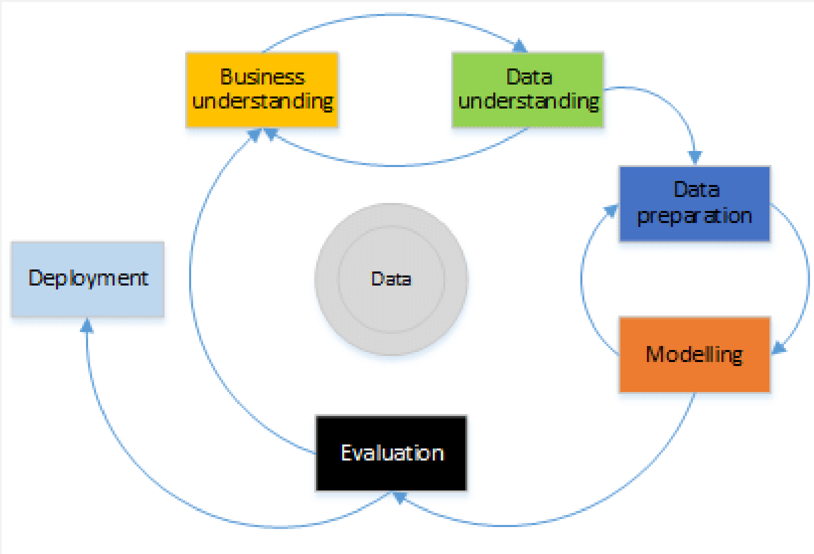
- **Employee Attrition Study** |Individual Report

rapidminer

Robin Roy
M Sc. Data Analytics

Project title:	Employee attrition study by data mining, using CRISP-DM methodology
Outcomes:	Studied the Likelihood & Indicators of attrition of employees from the company and strategies to reduce attrition are proposed as a summary studying the likelihood and indicators.
Project Description and Key Lessons-Learned	
Brief description of context	The project enabled to learn that data mining as the core process where a number of complex and intelligent methods are applied to extract patterns from data. Data mining process includes a number of tasks such as association, classification, prediction, clustering, time series analysis and so on. Here a fictional dataset provided by IBM data scientists is used to study the process of data mining using CRISP-DM. The attrition in other words is a natural process of loss of employees from the company. It is studied understanding how to explore, clean data, create different models, evaluating them to find the best model and its deployment process using rapid miner.
State holders	They are the interest groups which we need to consider in the process mining. They are the employees whose attrition is studied. The second interest group involved is the Human Resource department of the company and the top lop level administrators. There are also shareholders as important stakeholder. The employee stakeholder supports the current analysis
Key metrics of success	Metrics are important to take the right decision as like as choosing the right data mining technique. **Timely:** Process should be completed in the time frame where time is an important metric since the insights mined might not be useful after a particular time, for example in this case where most of the employees left the company. **Relevancy:** The mined analytics should be relevant in problem addressed. **Accuracy:** Accuracy is a measure of how well the model correlates an outcome with the attributes in the data that has been provided. **Reliability:** Reliability assesses the way that a data mining model performs on different data sets. A data mining model is reliable if it generates the same type of predictions or finds the same general kinds of patterns regardless of the test data that is supplied. **Usefulness:** Usefulness includes various metrics that tell us whether the model provides useful information. The data mining model that correlates

	employee attrition might be both accurate and reliable, but might not be useful in all scenarios.
Methodology	The CRISP-DM methodology and its usefulness is learnt. CRISP enhances the process of datamining. It is a 6-step process, and the key points understood in (1) *Business Understanding,* we should understand the business goals where we try to find stake holders, business experts and domain experts. We run operators to find; statistics: mean and median; imperative statistics: how similar the two data set of data are, regression statistics: relation between statistics and analyses all the relations, correlation to understanding relatability of two variables. In (2,3) Data understanding/exploration is organized by source, acquisition method, and potential errors, then visualized for further review. The most useful data is selected, cleaned, and integrated across multiple databases. Missing values are replaced either by the mean or mode of the variables and duplicates are removed. (4) In *Modelling* adequate techniques chosen, the data models selected are built and tested. Studied that in classification model we try to predict categorical models. In prediction models we try to predict the numerical or the continuous variable. In (5) *Evaluation* the data model is reviewed for utility, completeness, and ability to meet established business requirements. After finding the model accuracies, we decide which model to use. During (6) Deployment we have to put the created models into action to realize their full value. The main purpose of a deployment is to take new data as input which is known as the score data and return results.



i)Business understanding:

The current study studies the attrition of employees from the company using the fictional open source dataset created by IBM data scientists. Employee attrition is sometimes natural but when the attrition rate starts to grow past some certain threshold, it is an area of concern. Attrition may lead to overhead charges in training and recruiting new staffs while compromising outputs. Employee attrition can differ among organizations based on the kind of people leaving but the definition of attrition remains the same.

Business problem statement

When employees resign from the company, costs are incurred in recruiting new employees and training them. Productivity will be lower until new hires learn the business. If the new hire is not proficient the company could lose clients who are dissatisfied with service decreasing the revenue. Therefore, the primary business objective of the project is to retain the current employees by predicting the probability of leaving. Hence, we need to

- Create a model with the most accuracy in the prediction of attrition with current data to use the model with new datasets.
- If the new datasets vary in number of factors, changes to the algorithm can be made using drifts.
- Identify the variables that contribute maximum to the attrition.
- Create strategies to reduce attrition.

Analytics Problem Statement and its suitability

Problem statement as an analytics problem with constraints:
The deliverable of the data mining is what are the reasons for the attrition of employees from the company, the question is smart enough to deliver impact to the company. The

primary business objective of the project is to keep the current employee retention by predicting the probability of their leaving.

There are many other commercial issues that are related to the problem:

- The company identifies a business concern to be satisfied.
- Sources of potential raw data, and their sources, are identified.
- The data model is built based on the available data.
- The data structure, based on the data model, is built.
- The data structure is mined for useful information, interesting patterns, etc.

The company can only consider improving the availability of resource to employees and not inorganic factors for maximum employee retention.

ii)Data Understanding

The open dataset selected is the HR Analytics Employee Attrition & Performance of IBM which is collected by their CRM software, the software records different parameters of an employee of the company (such as satisfaction level, Salary, number of promotions, left the company etc.) The dataset is best used for the prediction of the attrition of the company's valuable employees.

Data source: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

We analyse the common grounds in employee attrition.

```
import io
attrition_data = pd.read_csv(io.BytesIO(uploaded['HR-Employee-Attrition.csv']))
attrition_data.head()
```

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	EnvironmentSatis
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	

```
attrition_data.columns
```

```
Index(['Age', 'Attrition', 'BusinessTravel', 'DailyRate', 'Department',
       'DistanceFromHome', 'Education', 'EducationField', 'EmployeeCount',
       'EmployeeNumber', 'EnvironmentSatisfaction', 'Gender', 'HourlyRate',
       'JobInvolvement', 'JobLevel', 'JobRole', 'JobSatisfaction',
       'MaritalStatus', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked',
       'Over18', 'OverTime', 'PercentSalaryHike', 'PerformanceRating',
       'RelationshipSatisfaction', 'StandardHours', 'StockOptionLevel',
       'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance',
       'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion',
       'YearsWithCurrManager'],
      dtype='object')
```

```
attrition_data.info
```

```
<bound method DataFrame.info of       Age Attrition  ... YearsSinceLastPromotion  YearsWithCurrManager
0      41       Yes  ...                       0                     5
1      49        No  ...                       1                     7
2      37       Yes  ...                       0                     0
3      33        No  ...                       3                     0
4      27        No  ...                       2                     2
...   ...       ...  ...                     ...                   ...
1465   36        No  ...                       0                     3
1466   39        No  ...                       1                     7
1467   27        No  ...                       0                     3
1468   49        No  ...                       0                     8
1469   34        No  ...                       1                     2

[1470 rows x 35 columns]>
```



```
Data types: int64 (26), object (9)
Range index: 1470 entries, 0 to 1469
Data columns (total 35 columns)
```

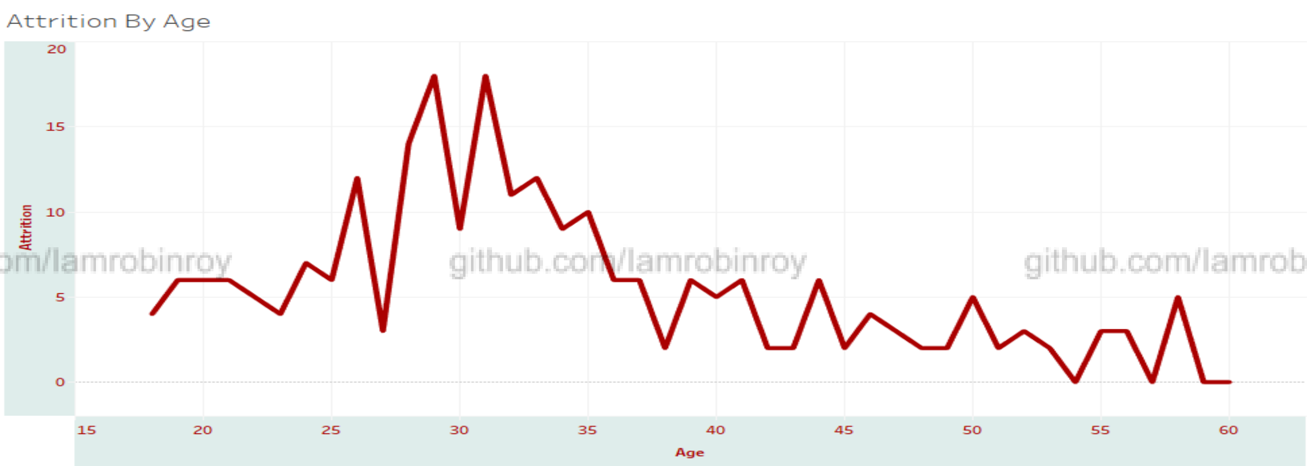
Employee number is a unique identifier

```
attrition_data.dtypes #finding the datatypes int64,object,float64
```

```
Age                          int64 PerformanceRating          int64
Attrition                   object RelationshipSatisfaction   int64
BusinessTravel              object StandardHours              int64
DailyRate                    int64 StockOptionLevel           int64
Department                  object TotalWorkingYears          int64
DistanceFromHome             int64 TrainingTimesLastYear      int64
Education                    int64 WorkLifeBalance            int64
EducationField              object YearsAtCompany             int64
EmployeeCount                int64 YearsInCurrentRole         int64
EmployeeNumber               int64 YearsSinceLastPromotion    int64
EnvironmentSatisfaction      int64 YearsWithCurrManager       int64
Gender                      object dtype: object
HourlyRate                   int64
JobInvolvement               int64
JobLevel                     int64
JobRole                     object
JobSatisfaction              int64
MaritalStatus               object
MonthlyIncome                int64
MonthlyRate                  int64
NumCompaniesWorked           int64
Over18                      object
OverTime                    object
PercentSalaryHike            int64
PerformanceRating            int64
```

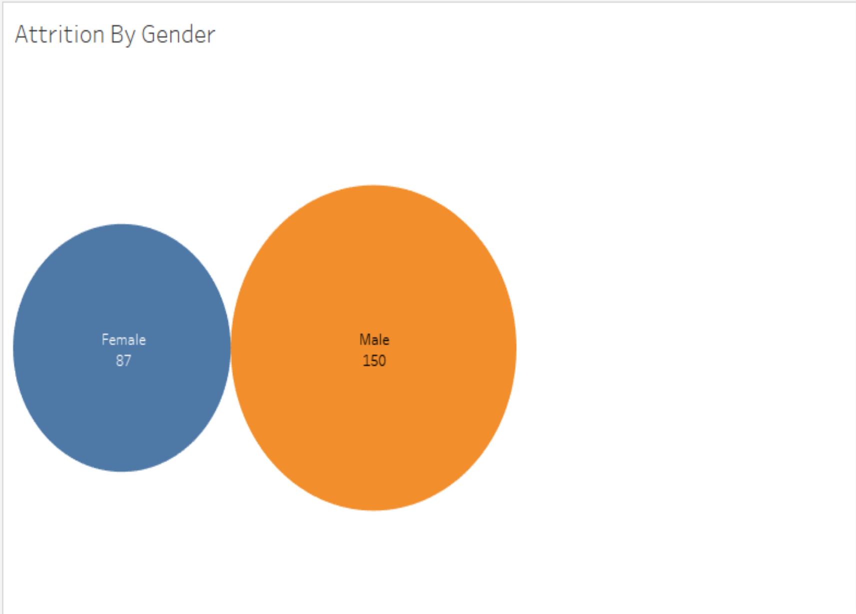
```
Most Positive Correlations:
 PerformanceRating     0.002889
MonthlyRate           0.015170
NumCompaniesWorked    0.043494
DistanceFromHome      0.077924
Target                1.000000
Name: Target, dtype: float64

Most Negative Correlations:
 TotalWorkingYears    -0.171063
JobLevel             -0.169105
YearsInCurrentRole   -0.160545
MonthlyIncome        -0.159840
Age                  -0.159205
Name: Target, dtype: float64
```

some of the insights derived are discussed below:



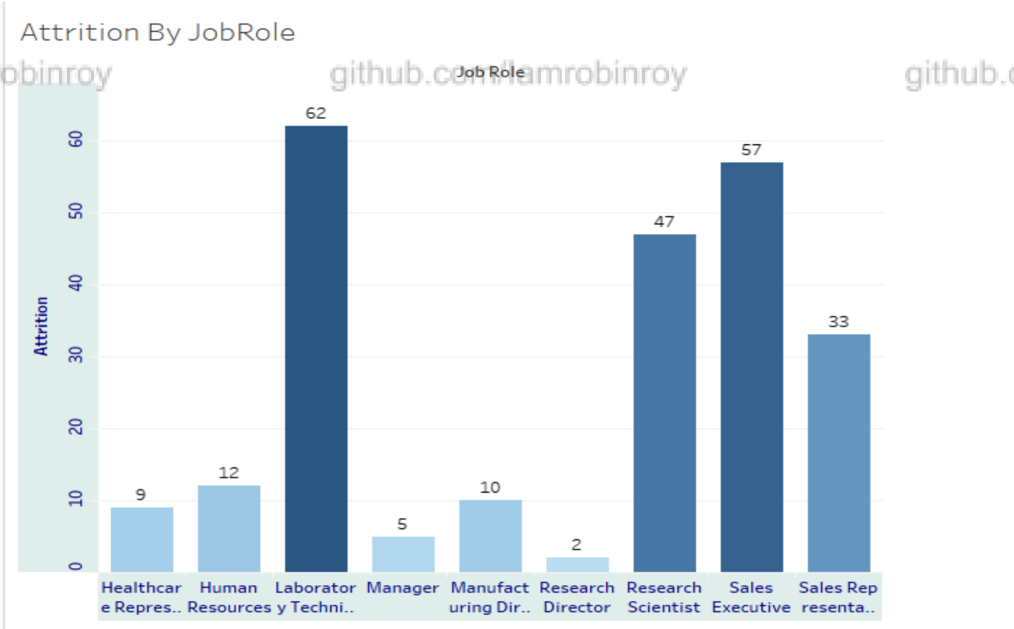
Attrition by age: The above fig. shows employee attrition in the company according to age. Staffs with age group 25 to 35 seems to have high level of attrition. The employees in big age category posses very less attrition compared to young.

Age distribution is little bit slightly right-skewed normal distribution with highest employees in the rage of 25 and 45 years old. The average years active employees stay is 7.37 years where it is 5.31 years for employees who already left.



z

Attrition by age: The gender Female are less likely to leave compared to the Males. After normalization the distribution of ex-employees is 17% for Males and 14.8% for Females.



Attrition by job role: Laboratory technician and sales executive represents maximum attrition, where very less in case of managers and directors where the probability of their attrition is very less.

Software/tools used

Rapid Miner: Rapid miner is a tool created for data mining that support the analysts for the purpose. To make the data mining process more transparent and smoother, it has a good set of predefined operators solving a wide range of problems.

Python: Python is the most popular programming language that offers the flexibility and power for programmers and data scientists to perform data analysis and apply machine learning algorithms. Python has become more popular for data mining due to the rise in the number of data analysis libraries.

Tableau: Tableau is a widely used resource for data visualization and business intelligence, and is focused on enhancing the analytic workflow experience.

iii)Data Preparation

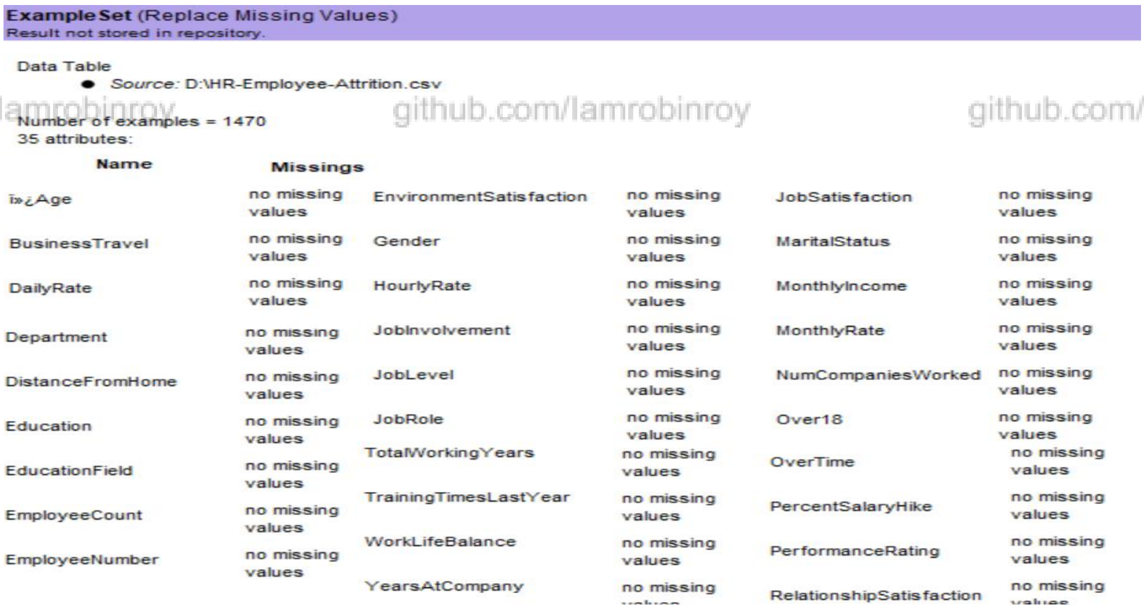
<u>Harmonize, rescale and clean data</u>

The data in the real world is always incomplete, noisy, and inconsistent because of not applicable, human or computer error at data entry, errors in data transmission, or from different data sources, etc. Therefore, the major tasks in data pre-processing includes data cleaning, data integration, data transformation, and data reduction.

Data Cleaning removes inaccurate, incomplete data from the source. Data is cleaned by either restoring missing data or removing the noisy data.

<u>Replacing missing values with zeros</u>

Missing data can be added manually, replaced with a calculated mean or average, or simply replaced with the most probable value as calculated by the team.



ExampleSet (Replace Missing Values)
Result not stored in repository.

Data Table

- *Source:* D:\HR-Employee-Attrition.csv

Number of examples = 1470
35 attributes:

Name	Missings				
ï»¿Age	no missing values	EnvironmentSatisfaction	no missing values	JobSatisfaction	no missing values
BusinessTravel	no missing values	Gender	no missing values	MaritalStatus	no missing values
DailyRate	no missing values	HourlyRate	no missing values	MonthlyIncome	no missing values
Department	no missing values	JobInvolvement	no missing values	MonthlyRate	no missing values
DistanceFromHome	no missing values	JobLevel	no missing values	NumCompaniesWorked	no missing values
Education	no missing values	JobRole	no missing values	Over18	no missing values
EducationField	no missing values	TotalWorkingYears	no missing values	OverTime	no missing values
EmployeeCount	no missing values	TrainingTimesLastYear	no missing values	PercentSalaryHike	no missing values
EmployeeNumber	no missing values	WorkLifeBalance	no missing values	PerformanceRating	no missing values
		YearsAtCompany	no missing	RelationshipSatisfaction	no missing

<u>Removing duplicates</u>

ExampleSet (1,470 examples, 1 special attribute, 34 regular attributes)

Row No.	ï»¿Age	DailyRate	DistanceFro...	Education	EmployeeCo...	EmployeeNu...	Environment...	HourlyRate	JobInvolvem...
1	0.446	0.742	-1.011	-0.891	0	-1.701	-0.660	1.383	0.380
2	1.322	-1.297	-0.147	-1.868	0	-1.699	0.255	-0.241	-1.026
3	0.008	1.414	-0.887	-0.891	0	-1.696	1.169	1.284	-1.026
4	-0.430	1.461	-0.764	1.061	0	-1.694	1.169	-0.487	0.380
5	-1.086	-0.524	-0.887	-1.868	0	-1.691	-1.575	-1.274	0.380
6	-0.539	0.502	-0.887	-0.891	0	-1.689	1.169	0.645	0.380
7	2.417	1.292	-0.764	0.085	0	-1.686	0.255	0.743	1.785
8	-0.758	1.377	1.827	-1.868	0	-1.684	1.169	0.055	0.380
9	0.118	-1.453	1.703	0.085	0	-1.682	1.169	-1.077	-1.026
10	-0.101	1.230	2.197	0.085	0	-1.681	0.255	1.383	0.380
11	-0.211	0.016	0.840	0.085	0	-1.679	-1.575	0.891	1.785
12	-0.867	-1.610	0.716	-0.891	0	-1.677	1.169	-0.831	-1.026
13	-0.648	-0.328	2.073	-1.868	0	-1.676	-1.575	-1.716	0.380

Past normalization we detect the outliers removing 10 biggest outliers from the dataset.

Result History | ExampleSet (Normalize) | ExampleSet (Detect Outlier (Distances))

Data | Statistics | Visualizations | Annotations

Open in Turbo Prep Auto Model

Filter (1,470 / 1,470 examples): all

Row No.	outlier ↓	Age	DailyRate	DistanceFro...	Education	Environment...	HourlyRate	JobInvolvem...	JobLevel	JobSatisfac...	MonthlyInco...
99	true	2.307	-0.299	0.100	1.061	1.169	-1.421	0.380	1.749	0.246	1.565
127	true	2.307	-1.624	1.703	1.061	1.169	1.383	0.380	0.846	1.153	0.809
179	true	0.994	-0.685	-1.011	-0.891	-0.660	1.284	0.380	0.846	-1.567	0.839
402	true	2.088	0.257	-0.394	0.085	0.255	0.989	1.785	1.749	-1.567	1.425
428	true	2.526	1.726	2.320	0.085	0.255	0.694	-1.026	0.846	-1.567	0.799
596	true	2.307	-1.280	-0.887	1.061	1.169	-1.716	0.380	2.652	-0.661	2.707
862	true	0.994	1.486	-0.887	0.085	0.255	0.153	0.380	1.749	-1.567	2.240
1079	true	0.775	-1.652	2.320	0.085	1.169	-1.667	0.380	1.749	-1.567	2.087
1136	true	0.994	-0.594	-1.011	1.061	1.169	-0.487	1.785	1.749	-1.567	2.350
1402	true	1.979	-1.520	2.073	1.061	0.255	0.251	1.785	2.652	-0.661	2.790
1	false	0.446	0.742	-1.011	-0.891	-0.660	1.383	0.380	-0.058	1.153	-0.108
2	false	1.322	-1.297	-0.147	-1.868	0.255	-0.241	-1.026	-0.058	-0.661	-0.292
3	false	0.008	1.414	-0.887	-0.891	1.169	1.284	-1.026	-0.961	0.246	-0.937
4	false	-0.430	1.461	-0.764	1.061	1.169	-0.487	0.380	-0.961	0.246	-0.763

ExampleSet (1,470 examples, 1 special attribute, 31 regular attributes)

github.com/lamrobinroy github.com/lamrobinroy github.com/lamrobinroy

Additional or u necessary data can affect the prediction results. Attributes such as standardHours, over18, EmployeeCount where contained same values hence, we removed them. After removing those unnecessary features as well, the dataset has 31 features.

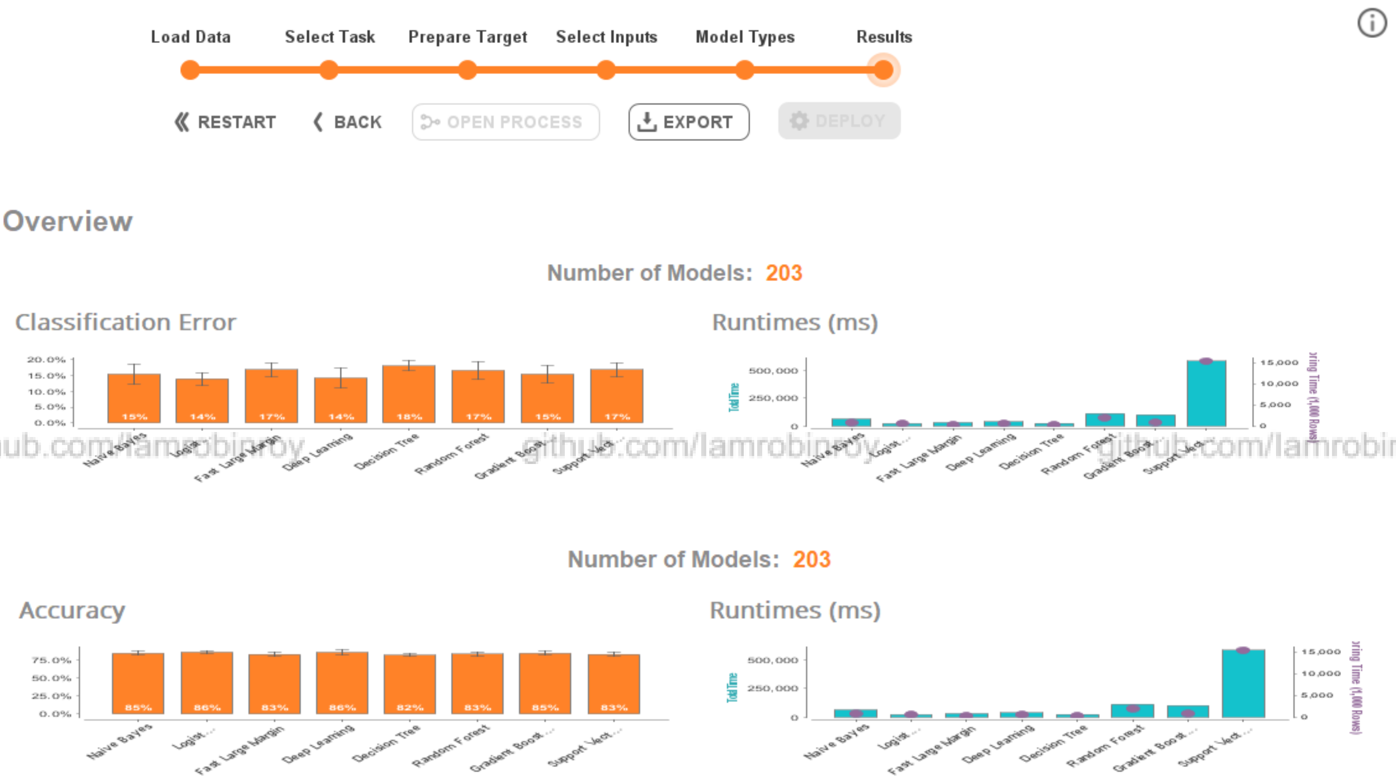
Name	Description
AGE	Numerical Value
ATTRITION	Employee leaving the company (0=no, 1=yes)
BUSINESS TRAVEL	(1=No Travel, 2=Travel Frequently, 3=Tavel Rarely)
DAILY RATE	Numerical Value - Salary Level
DEPARTMENT	(1=HR, 2=R&D, 3=Sales)
DISTANCE FROM HOME	Numerical Value - THE DISTANCE FROM WORK TO HOME
EDUCATION	Numerical Value

github.com/lamrobinroy github.com/lamrobinroy github.com/lamrobinroy

EDUCATION FIELD	(1=HR, 2=LIFE SCIENCES, 3=MARKETING, 4=MEDICAL SCIENCES, 5=OTHERS, 6= TEHCNICAL)
ENVIROMENT SATISFACTION	Numerical Value - SATISFACTION WITH THE ENVIROMENT
GENDER	(1=FEMALE, 2=MALE)
HOURLY RATE	Numerical Value - HOURLY SALARY
JOB INVOLVEMENT	Numerical Value - JOB INVOLVEMENT
JOB LEVEL	Numerical Value - LEVEL OF JOB
JOB ROLE	(1=HC REP, 2=HR, 3=LAB TECHNICIAN, 4=MANAGER, 5= MANAGING DIRECTOR, 6= REASEARCH DIRECTOR, 7= RESEARCH SCIENTIST, 8=SALES EXECUTIEVE, 9= SALES REPRESENTATIVE)
JOB SATISFACTION	Numerical Value - SATISFACTION WITH THE JOB
MARITAL STATUS	(1=DIVORCED, 2=MARRIED, 3=SINGLE)
MONTHLY INCOME	Numerical Value - MONTHLY SALARY
MONTHY RATE	Numerical Value - MONTHY RATE
NUMCOMPANIES WORKED	Numerical Value - NO. OF COMPANIES WORKED AT
OVERTIME	(1=NO, 2=YES)
PERCENT SALARY HIKE	Numerical Value - PERCENTAGE INCREASE IN SALARY. The parentage of change in salary between 2 year (2017, 2018).
PERFORMANCE RATING	Numerical Value - ERFORMANCE RATING
RELATIONS SATISFACTION	Numerical Value - RELATIONS SATISFACTION
STOCK OPTIONS LEVEL	Numerical Value - STOCK OPTIONS. How much company stocks you own from this company?
TOTAL WORKING YEARS	Numerical Value - TOTAL YEARS WORKED
TRAINING TIMES LAST YEAR	Numerical Value - HOURS SPENT TRAINING
WORK LIFE BALANCE	Numerical Value - TIME SPENT BEWTWEEN WORK AND OUTSIDE
YEARS AT COMPANY	Numerical Value - TOTAL NUMBER OF YEARS AT THE COMPNAY
YEARS IN CURRENT ROLE	Numerical Value -YEARS IN CURRENT ROLE
YEARS SINCE LAST PROMOTION	Numerical Value - LAST PROMOTION
YEARS WITH CURRENT MANAGER	Numerical Value - YEARS SPENT WITH CURRENT MANAGER

iv)Modelling:

HR department needs to overcome the problem of employee attrition. The objective of the modelling phase is to build multiple models and select the model giving maximum accuracy for predicting attrition. In different models built, performance, accuracy, class precision and class recall are calculated and compared on the test and trained dataset. As usual the dataset is divided into train set and test set datasets as 70% and 30% respectively. We use the auto model feature in rapid miner to find appropriate models before creating our models, it let us compare between accuracy, precision, classification between different models.



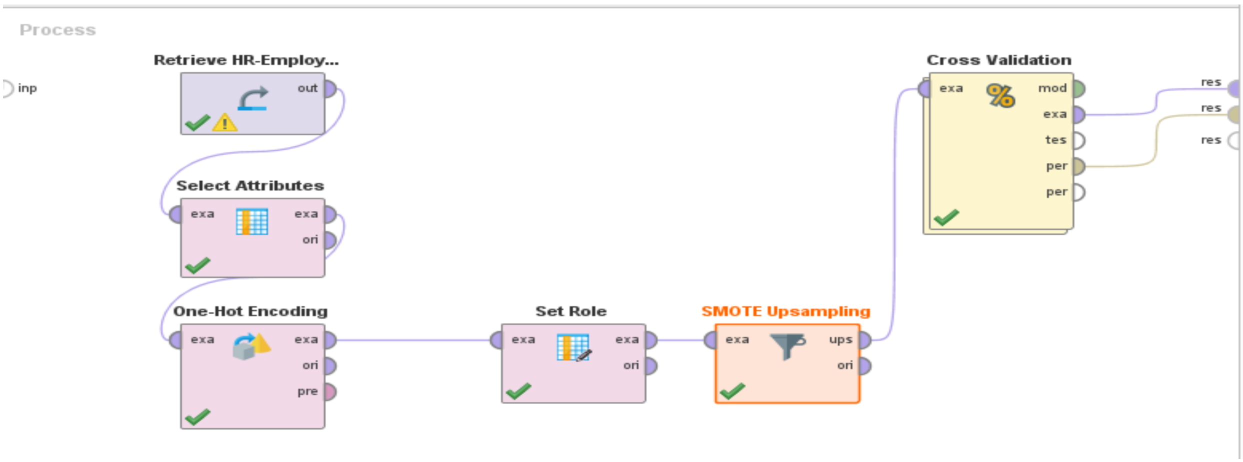
Auto Model generates a RapidMiner Studio process behind the scenes, helping us to fine tune and test models before putting them into production. This enable us to choose the best fitting model for our process.

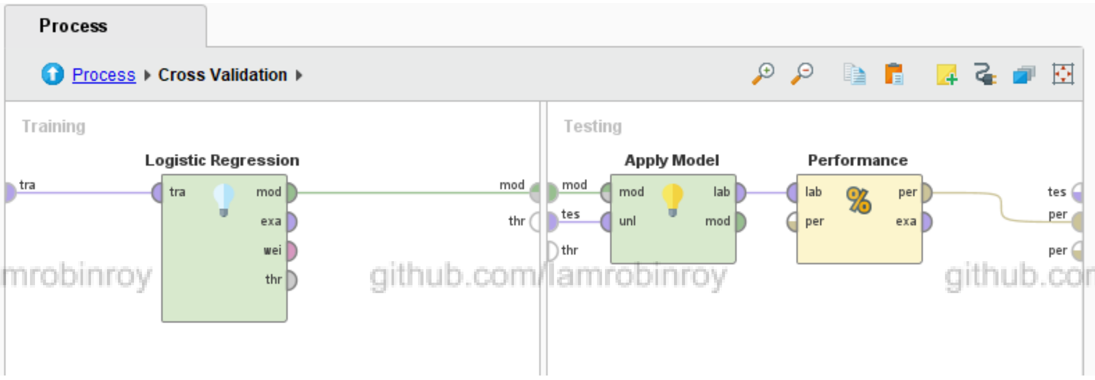
Logistic Regression

Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. In Logistic regression the target variable is set as 'attrition' which distinguishes the employees as active or not. We also set the target role to label.

e.g.- y = e^(b0 + b1*x) / (1 + e^(b0 + b1*x))

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$





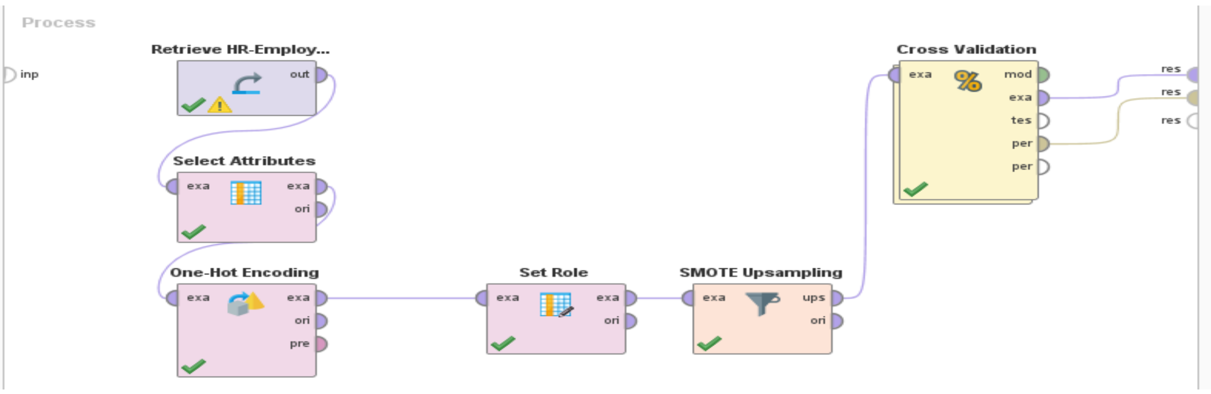
accuracy: 79.12% +/- 2.51% (micro average: 79.12%)

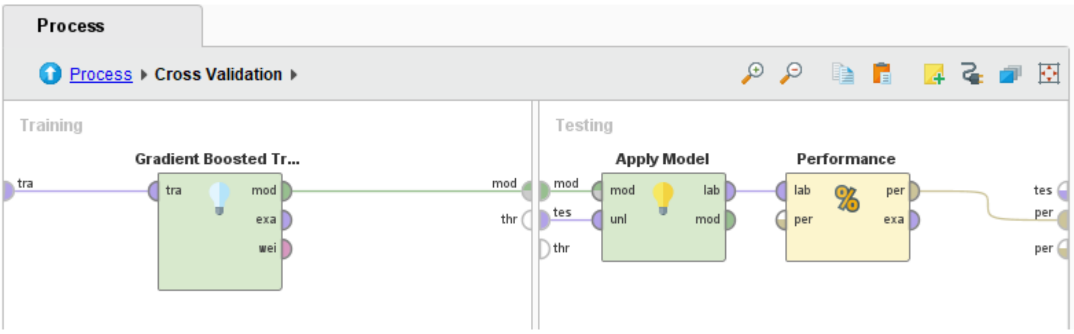
	true Yes	true No	class precision
pred. Yes	991	273	78.40%
pred. No	242	960	79.87%
class recall	80.37%	77.86%	

```
PerformanceVector:
accuracy: 79.12% +/- 2.51% (micro average: 79.12%)
ConfusionMatrix:
True:   Yes     No
Yes:    991     273
No:     242     960
classification_error: 20.88% +/- 2.51% (micro average: 20.88%)
ConfusionMatrix:
True:   Yes     No
Yes:    991     273
No:     242     960
AUC: 0.867 +/- 0.023 (micro average: 0.867) (positive class: No)
precision: 79.96% +/- 3.13% (micro average: 79.87%) (positive class: No)
ConfusionMatrix:
True:   Yes     No
Yes:    991     273
No:     242     960
recall: 77.86% +/- 3.89% (micro average: 77.86%) (positive class: No)
ConfusionMatrix:
True:   Yes     No
Yes:    991     273
No:     242     960
sensitivity: 77.86% +/- 3.89% (micro average: 77.86%) (positive class: No)
ConfusionMatrix:
True:   Yes     No
Yes:    991     273
No:     242     960
specificity: 80.37% +/- 3.82% (micro average: 80.37%) (positive class: No)
ConfusionMatrix:
True:   Yes     No
Yes:    991     273
No:     242     960
```


Gradient Boosted tree

We now use the gradient boosted tree to predict whether an employee would leave or stay at the organization. Gradient boosting trees builds tees in a serial manner, where each tree tries to correct the mistakes of the previous ones. the Gradient Boost trees have a depth larger than 1.





accuracy: 84.18% +/- 2.13% (micro average: 84.18%)

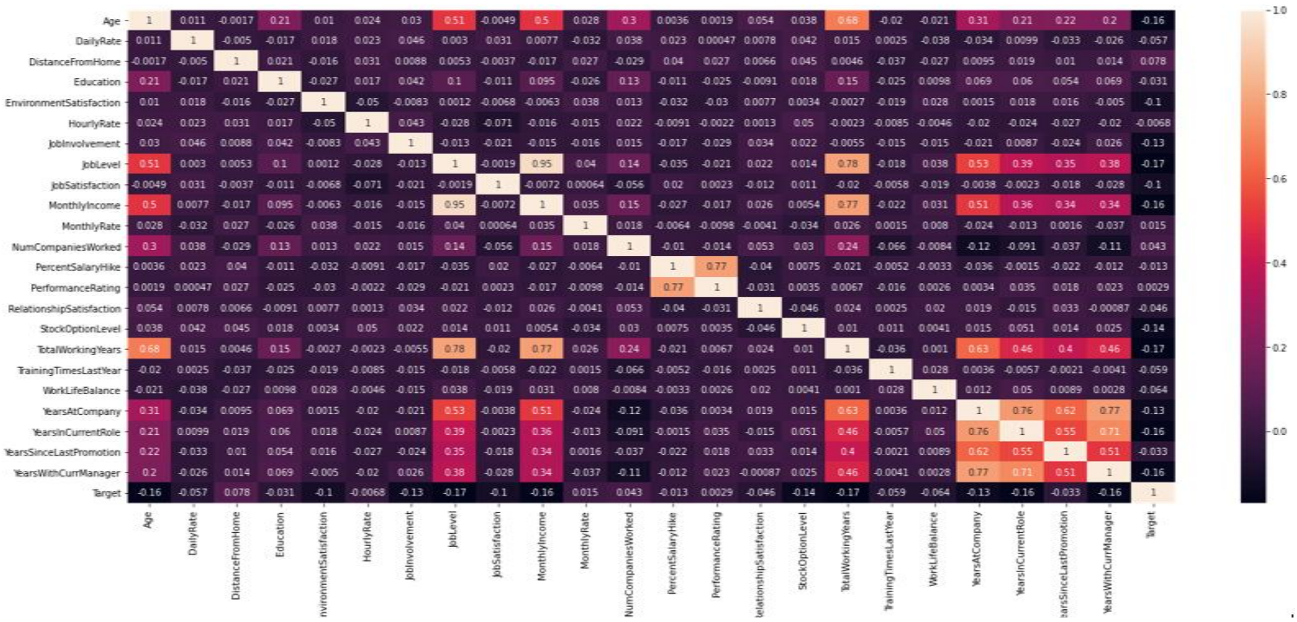
	true Yes	true No	class precision
pred. Yes	1023	180	85.04%
pred. No	210	1053	83.37%
class recall	82.97%	85.40%	

```
PerformanceVector:
accuracy: 84.18% +/- 2.13% (micro average: 84.18%)
ConfusionMatrix:
True:   Yes     No
Yes:    1023    180
No:     210     1053
classification_error: 15.82% +/- 2.13% (micro average: 15.82%)
ConfusionMatrix:
True:   Yes     No
Yes:    1023    180
No:     210     1053
precision: 83.47% +/- 2.84% (micro average: 83.37%) (positive class: No)
ConfusionMatrix:
True:   Yes     No
Yes:    1023    180
No:     210     1053
recall: 85.40% +/- 2.82% (micro average: 85.40%) (positive class: No)
ConfusionMatrix:
True:   Yes     No
Yes:    1023    180
No:     210     1053
sensitivity: 85.40% +/- 2.82% (micro average: 85.40%) (positive class: No)
ConfusionMatrix:
True:   Yes     No
Yes:    1023    180
No:     210     1053
specificity: 82.96% +/- 3.65% (micro average: 82.97%) (positive class: No)
ConfusionMatrix:
True:   Yes     No
Yes:    1023    180
No:     210     1053
```

Decision Tree

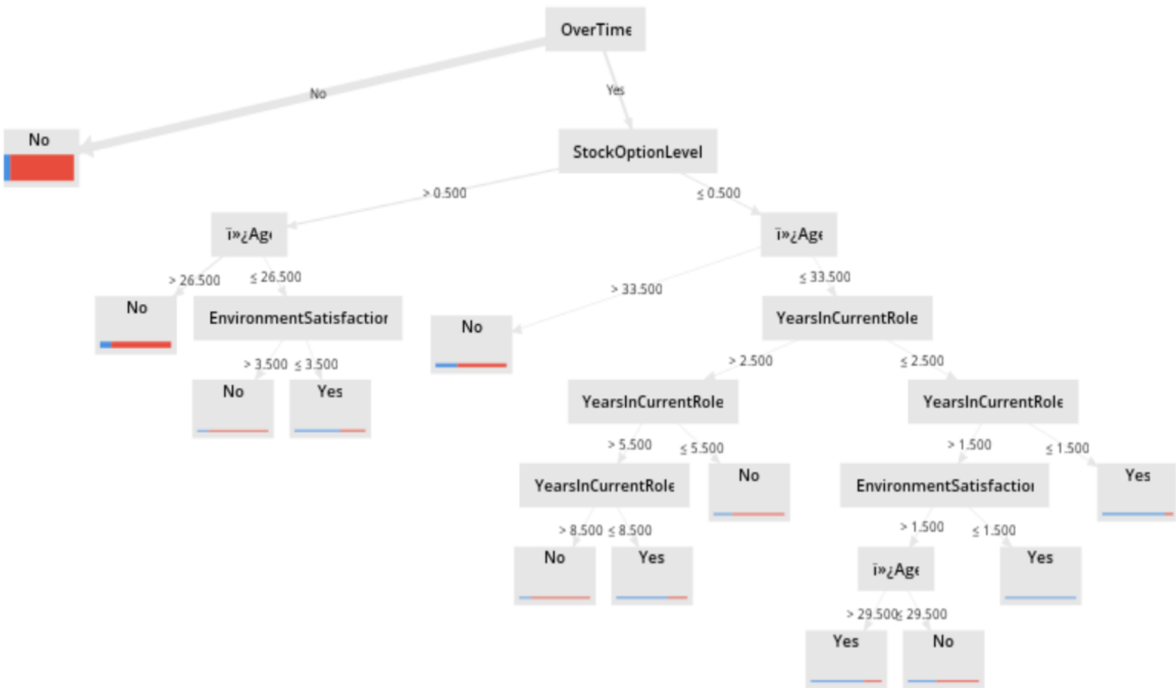
Decision trees, a type of algorithmic tool used to follow multiple potential paths to a desired goal and then identify the most effective one.

Correlation Matrix

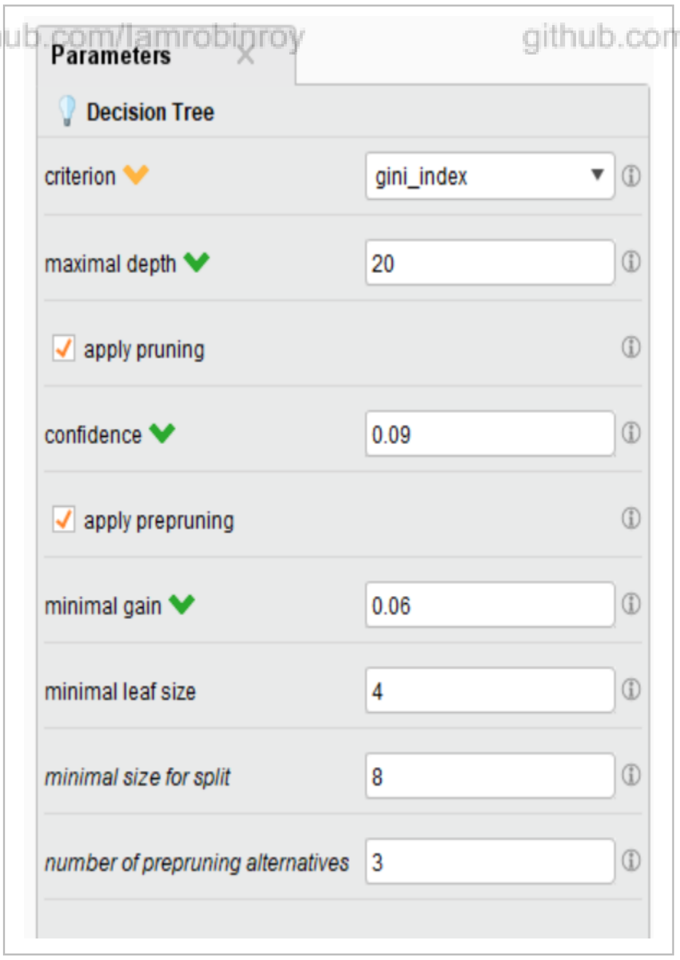


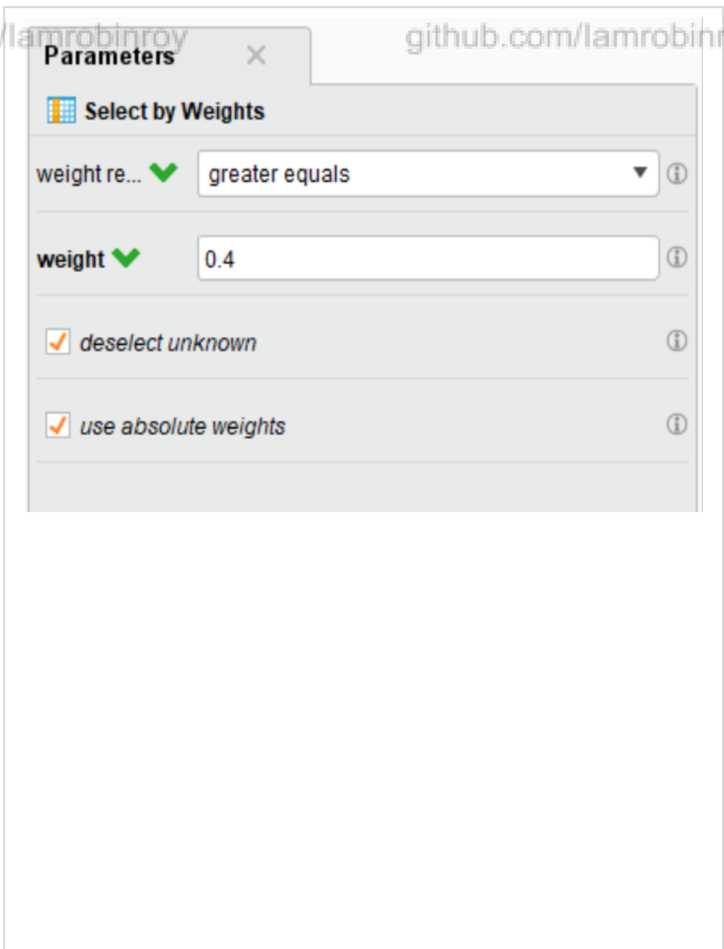
attribute	weight
ï»¿Age	0.593
EnvironmentSatisfaction	1
JobLevel	0
MonthlyIncome	0.033
OverTime	0.998
RelationshipSatisfaction	0.998
StockOptionLevel	0.999
YearsAtCompany	0.236
YearsInCurrentRole	0.434

Attribute Weights (Correlation Matrix)



Decision Tree





The decision tree shows that if they have overtime there are 944 who have not left the company and 110 people left the company.

If they have stock option which is grater that 0.500 and age greater than 26 there are just 33 people who have left the company. Checking the below path using the operator Get Decision Tree Path we are able to identify the individuals who left the company or didn't leave having the other attributes as path.

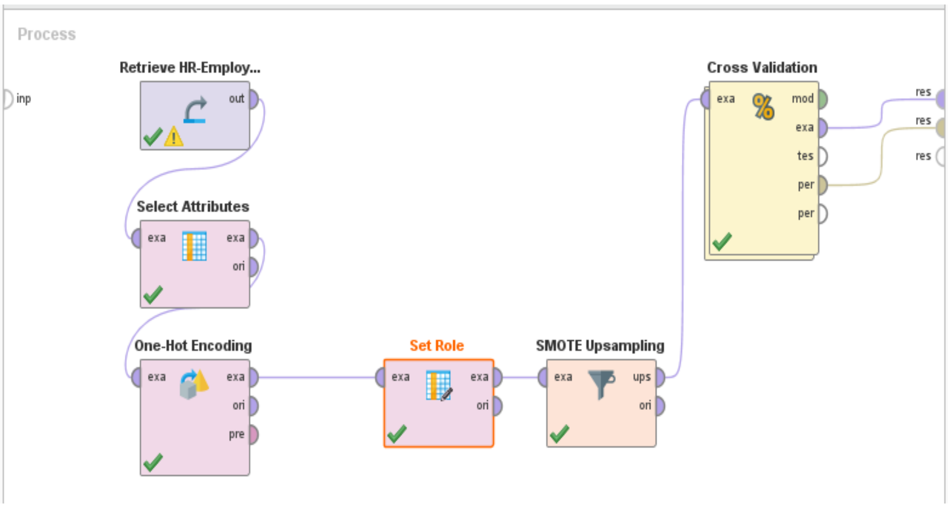
Row No.	Attrition	Path	ï»¿Age	EnvironmentSatisfaction	OverTime
232	No	OverTime = No	42	3	No
233	No	OverTime = No	59	2	No
234	No	OverTime = No	50	4	No
235	Yes	OverTime = Yes & StockOptionLevel > 0.500 & ï»¿Age > 26.500	33	3	Yes
236	No	OverTime = Yes & StockOptionLevel > 0.500 & ï»¿Age > 26.500	43	4	Yes
237	Yes	OverTime = No	33	1	No
238	No	OverTime = Yes & StockOptionLevel ≤ 0.500 & ï»¿Age > 33.500	52	1	Yes
239	No	OverTime = No	32	3	No
240	Yes	OverTime = Yes & StockOptionLevel ≤ 0.500 & ï»¿Age ≤ 33.500 & YearsInCurrentRole ≤ 2.500 & ...	32	4	Yes
241	No	OverTime = No	39	3	No
242	No	OverTime = No	32	3	No
243	No	OverTime = No	41	3	No

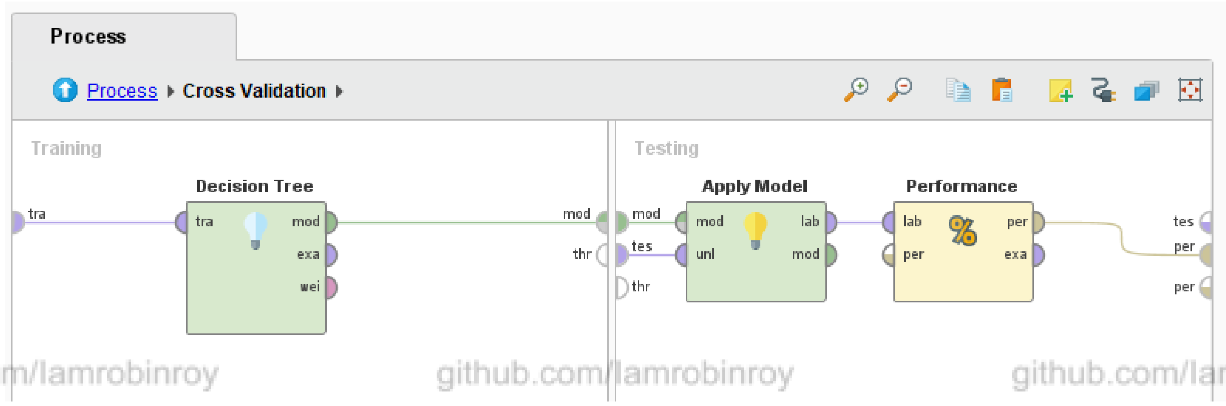
The pattern evaluation identifies truly interesting patterns representing knowledge based on different types of interestingness measures. A pattern is considered to be interesting if it is potentially useful, easily understandable by humans, validates some hypothesis that someone wants to confirm or valid on new data with some degree of certainty.

For. E.g. The employee number didn't leave the company even without the overtime with the age 42. The employee number 235 left the company after having over time and stock option greater than 0.500 and age is 26.

Row No.	Attrition ↓	Path
1397	Yes	OverTime = Yes & StockOptionLevel ≤ 0.500 & ï»¿Age > 33.500

People who fit the above path are more likely to leave than others





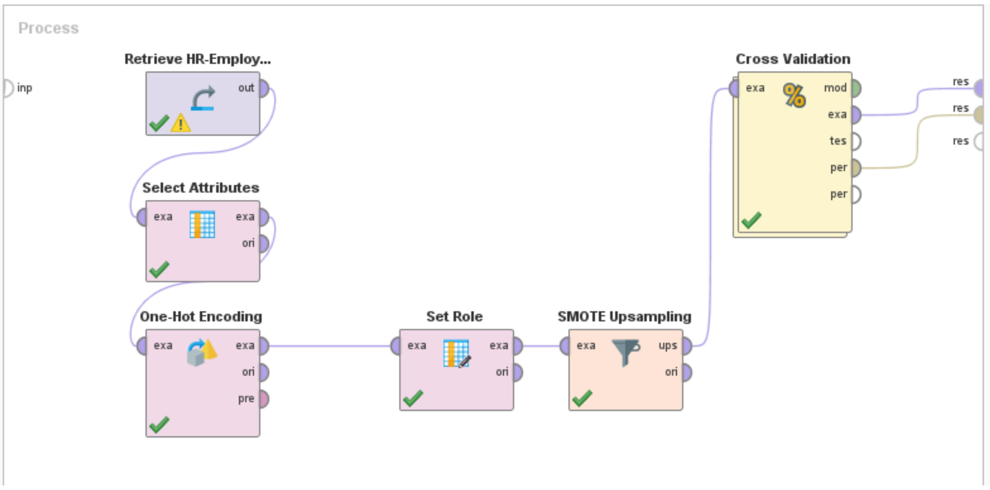
accuracy: 82.16% +/- 2.02% (micro average: 82.16%)

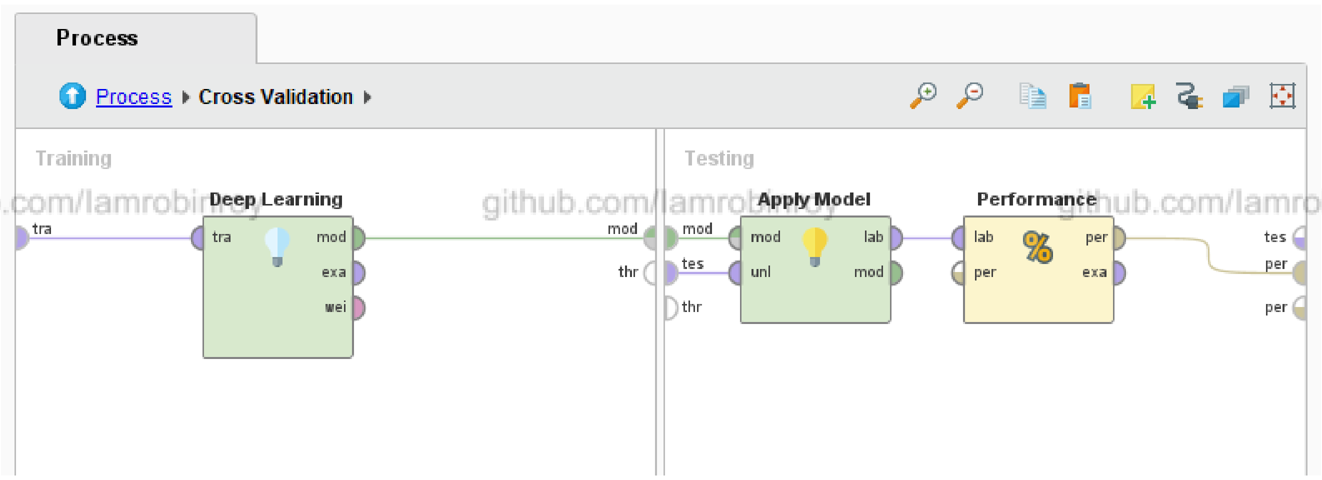
	true Yes	true No	class precision
pred. Yes	1022	229	81.69%
pred. No	211	1004	82.63%
class recall	82.89%	81.43%	

```
PerformanceVector:
accuracy: 82.16% +/- 2.02% (micro average: 82.16%)
ConfusionMatrix:
True:   Yes     No
Yes:    1022    229
No:     211     1004
classification_error: 17.84% +/- 2.02% (micro average: 17.84%)
ConfusionMatrix:
True:   Yes     No
Yes:    1022    229
No:     211     1004
AUC: 0.850 +/- 0.021 (micro average: 0.850) (positive class: No)
precision: 82.63% +/- 1.31% (micro average: 82.63%) (positive class: No)
ConfusionMatrix:
True:   Yes     No
Yes:    1022    229
No:     211     1004
recall: 81.43% +/- 4.05% (micro average: 81.43%) (positive class: No)
ConfusionMatrix:
True:   Yes     No
Yes:    1022    229
No:     211     1004
sensitivity: 81.43% +/- 4.05% (micro average: 81.43%) (positive class: No)
ConfusionMatrix:
True:   Yes     No
Yes:    1022    229
No:     211     1004
specificity: 82.89% +/- 1.50% (micro average: 82.89%) (positive class: No)
ConfusionMatrix:
True:   Yes     No
Yes:    1022    229
```

Deep learning

Deep learning algorithm run data through several “layers” of neural network algorithms, each of which passes a simplified representation of the data to the next layer.



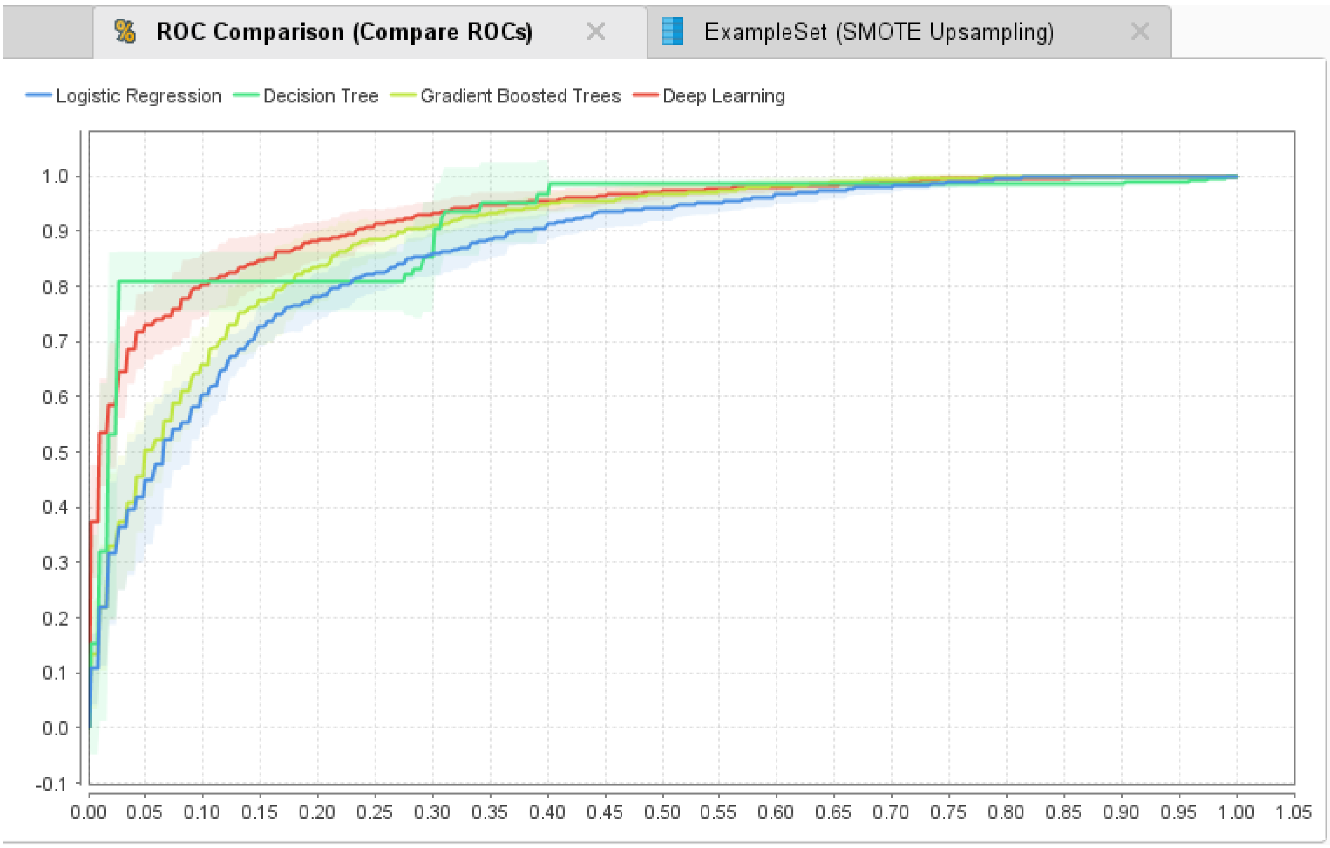


accuracy: 84.75% +/- 2.03% (micro average: 84.75%)

	true Yes	true No	class precision
pred. Yes	1050	193	84.47%
pred. No	183	1040	85.04%
class recall	85.16%	84.35%	

```
PerformanceVector:
accuracy: 84.75% +/- 1.68% (micro average: 84.75%)
ConfusionMatrix:
True:   Yes     No
Yes:    1059    202
No:     174     1031
classification_error: 15.25% +/- 1.68% (micro average: 15.25%)
ConfusionMatrix:
True:   Yes     No
Yes:    1059    202
No:     174     1031
AUC: 0.928 +/- 0.014 (micro average: 0.928) (positive class: No)
precision: 85.68% +/- 2.68% (micro average: 85.56%) (positive class: No)
ConfusionMatrix:
True:   Yes     No
Yes:    1059    202
No:     174     1031
recall: 83.61% +/- 3.94% (micro average: 83.62%) (positive class: No)
ConfusionMatrix:
True:   Yes     No
Yes:    1059    202
No:     174     1031
sensitivity: 83.61% +/- 3.94% (micro average: 83.62%) (positive class: No)
ConfusionMatrix:
True:   Yes     No
Yes:    1059    202
No:     174     1031
specificity: 85.88% +/- 3.26% (micro average: 85.89%) (positive class: No)
ConfusionMatrix:
True:   Yes     No
Yes:    1059    202
No:     174     1031
```


ROC Curve



We use ROC (Receiver Operating Characteristics) When we need to check or visualize the performance of the multi - class classification problem. ROC is a probability curve. This is an ideal situation. When two curves don't overlap at all means model has an ideal measure of separability. When two distributions overlap, we introduce type 1 and type 2 error. Depending upon the threshold, we can minimize or maximize them

v)Evaluation:

The table below lists the accuracy, precision, recall, AUC, sensitivity and specify of our models; deep learning, gradient boosted tree, decision tree and logistic regression. We compare them to select the best suitable model in terms of accuracy and precision for the deployment.

Models	Accuracy (%)	Precision (%)	Recall (%)	AUC (%)	Sensitivity (%)	Specificity (%)
Deep Learning	84.75 +/- 1.68	85.68 +/- 2.68	83.61 +/- 3.94	0.968 +/- 0.014	83.61 +/- 3.94	85.88 +/- 3.26
Gradient Boosted Tree	84.18 +/- 2.13	83.47 +/- 2.84	85.40 +/- 2.82	0.940 +/- 0.021	85.40 +/- 2.82	82.96 +/- 3.65
Decision Tree	82.16 +/- 2.02	82.63 +/- 1.31	81.43 +/- 4.05	0.850 +/- 0.021	81.43 +/- 4.05	82.89 +/- 1.50
Logistic Regression	79.12 +/- 2.51	79.96 +/- 3.13	77.86 +/- 3.89	0.867 +/- 0.023	77.86 +/- 3.89	80.37 +/- 3.82

In the evaluation step we compared the models for employee attrition and found out that the model Deep learning gives better results than other models when the accuracy and precision is compared.

Deployment:

To realize the full value of the models, it is important to put them into production. A deployment plan is created which included processes used to monitor data mining for utility and accuracy.

Deployment location	Folder Location
Local	RapidMiner Studio repository
Remote	RapidMiner Server repository

The deployment in rapid miner studio repository can be shared and controlled by configuring user access. Then we select the deployment option to create new deployment location.

The next step is to activate monitoring, we select a pre-existing "PostgreSQL" as connection and we need to set up alerts like E-mail alerts, after our location set up is ready, we Create the Location. Now our location is ready to add deployments, we load our dataset

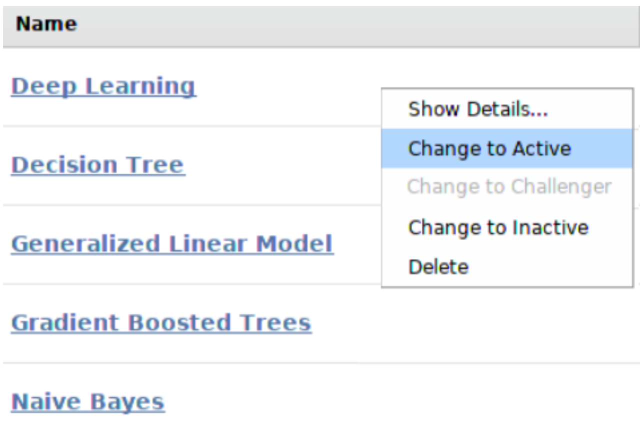
Local repository > Employee Attrition

We create a deployment called "Attrition", and identify our problem as a classification problem, before proceeding to build the models using Auto Model. The issue here is not merely to predict which employee of the company will leave, but to calculate the gains achievable by the model if it can correctly identify the churners.

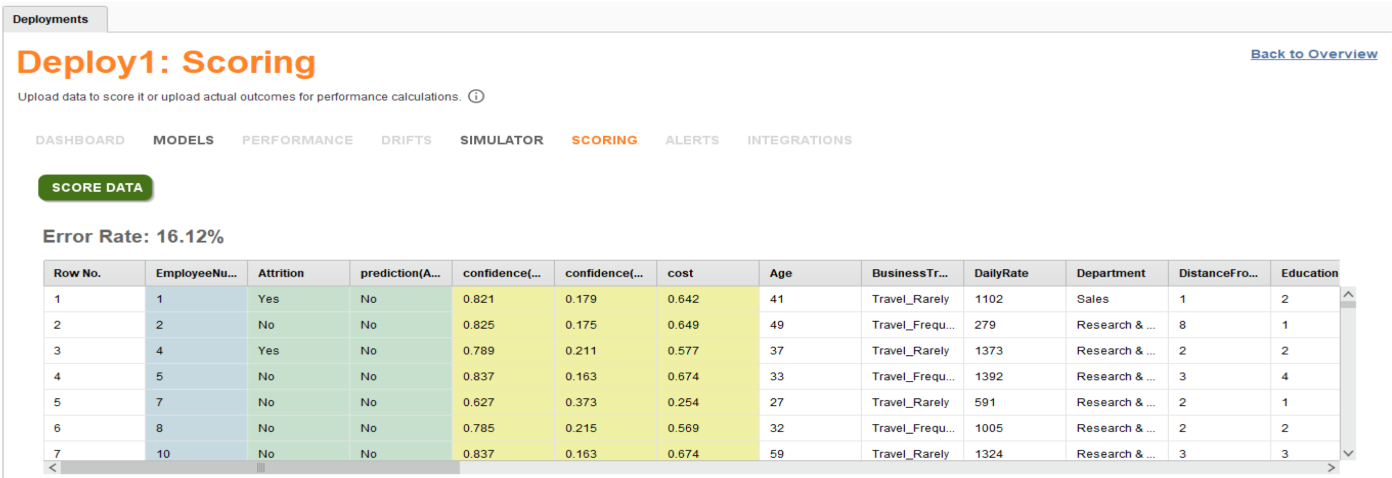
Deploy the model Deploying the model consists of three steps.

- Name the model ("Deep learning")
- Select the deployment location ("Remote_Deployments")
- Select the deployment folder (e.g., "Attrition")

Usually, a deployment will contain multiple models. But we use the model with the best performance. Since **Deep learning** has better performance than, we right-click the model and select **Change to Active**.



The main purpose of a deployment is to score data, in other words to use it with new data. The models take new data as input, and return a result. We use the **Score Data** function for this purpose, and choose a data set from the repository. The selected data is supposed to have similar columns which we used to build the models, and any extra columns will be removed, any values in the new dataset columns are missing they will be added by the mean values or the mode.



Row No.	EmployeeNu...	Attrition	prediction(A...	confidence(...	confidence(...	cost	Age	BusinessTr...	DailyRate	Department	DistanceFro...	Education
1	1	Yes	No	0.821	0.179	0.642	41	Travel_Rarely	1102	Sales	1	2
2	2	No	No	0.825	0.175	0.649	49	Travel_Frequ...	279	Research & ...	8	1
3	4	Yes	No	0.789	0.211	0.577	37	Travel_Rarely	1373	Research & ...	2	2
4	5	No	No	0.837	0.163	0.674	33	Travel_Frequ...	1392	Research & ...	3	4
5	7	No	No	0.627	0.373	0.254	27	Travel_Rarely	591	Research & ...	2	1
6	8	No	No	0.785	0.215	0.569	32	Travel_Frequ...	1005	Research & ...	2	2
7	10	No	No	0.837	0.163	0.674	59	Travel_Rarely	1324	Research & ...	3	3

In the scoring data only columns that are not used by the model as input can be identified as ID or target columns, we can use prediction to find the target values. We can resubmit the data with the ID later which we used to identify the data, when the target values are known, to generate error rates and other statistics.

After we apply and run the model the scoring data is color-coded to indicate its importance for the prediction

- Dark green values: Strongly support the prediction for that row of data.
- Dark red values: Strongly support a *different* prediction for that row of data
- Lighter colours: Less important.

The Dashboard provides the following statistics, displayed over time. We can choose the time interval and for more details we can see the Performance summary.

We can later use different web services for integration of the model into softwares by providing connection to the location of the model in rapid miner.

Conclusion

As datasets with similar structure accumulated from the company the algorithm can be re-trained using the additional data in order to generate more accurate predictions to identify employees with high attrition risk

We can assign a "**Attrition Score**" based on the prediction where;

- Less-attrition-risk for employees with label < 0.6
- Moderate-attrition-risk for employees with label between 0.6 and 0.8
- High-attrition-risk for employees with label > 0.8

The stronger indicators of people leaving include:

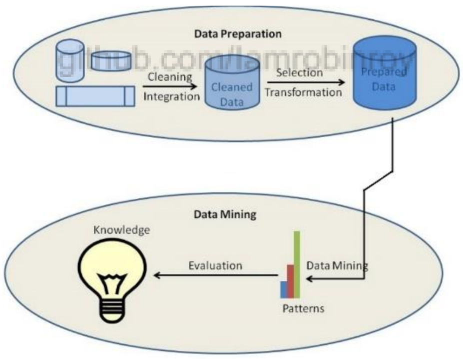
Monthly Income: people on higher salary are less likely to leave the company. It should be checked periodically if the company is providing competitive salaries according to the industry standards.

Over Time: People who work on over time are more likely to have attrition since actions must be taken in order to take measures to manage manpower according to the projects hence reducing over time, most overtime duties are not paid.

Age: Employees of age 25-35 are more likely to leave the company. Their reason might be that they are not given correct paths of promotions and incentives. Long-term vision should be given to employees to maintain the group of young employees.

DistanceFromHome: Employees who are living far are likely to leave the company, hence transportation facilities must be given if it’s a feasible option. Another option is selection of employees based on their location initially during interviews but its not a suggested option as long as the employee is ready to make it to work every day.

TotalWorkingYears: The most experienced employees are less likely to leave. Attrition of employees with less experience are more likely to leave. It can be seen as a reason of less commitment towards the company, it can be improved by taking actions by the HR team.



Data mining process using CRIS-DM methodology was very efficient in our current study, we were able to reach the above-mentioned conclusions and the deployment enables to use new score data for predictions and analysis.