SUBMITTED TO: TERRI HOARE
SUBMITTED BY: KOUSHIK CHIKKEGOWDA
STUDENT ID: 10515678

# Contents

# Abstract

Employee attrition is a term used when the strength of a company workforce reduces over a period due to an unavoidable situation like employee resignation for personal or other reasons. This situation is very problematic for the company because the company is losing more employees than they are hired. For example, few salespersons from a company could be asked to relocate to another branch and they may choose to resign for this reason. Some of the other reasons for the employee to leave is lack of professional development, bad working conditions, a job offer from another company. This type of loss can be detrimental for the company both in terms of productivity and financially. In this project, we are using IBM Employee attrition dataset where we will use rapid miner tool to perform machine learning to predict if an employee will leave the company or not. This can help the company to take preventive measures to avoid attrition from happening.

# Introduction

In the current job market employee jump from company to company due to various factors like job opportunities, Job stress level, Personal reasons. In big companies, if the employee is having a critical role then it very important to prevent attrition for those employees. Big companies may have thousands of employees within them and IBM is one such company. In order to help the company, we are using "Data Mining and Machine Learning Techniques" in rapid miner tool to predict and analyze attrition with IBM. The dataset we have chosen is from Kaggle which is an open source data repository.

The process was implemented using CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. CRISP-DM is about a non-proprietary; chronically open-source software, developed by field experts for performing better and faster results for data mining (Shearer, 2000). Every step in this assignment is implemented using CRISP-DM as out reference Model.

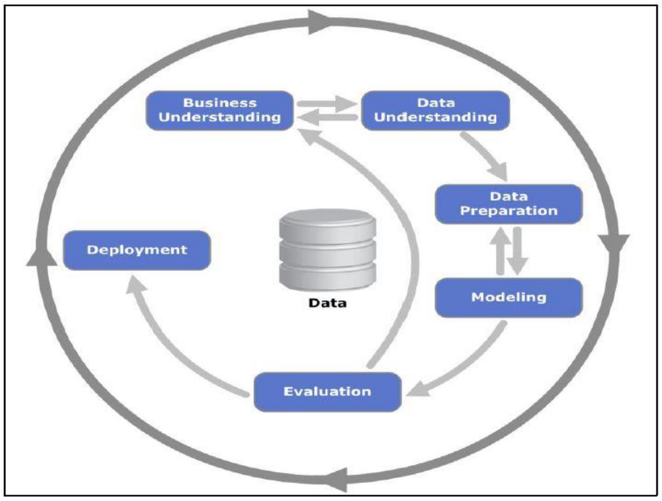


Figure 1: CRISP-DM Reference Model [Shearer, 2000]

### Business Understanding

In this section, we give an overview problem associated in our project and machine learning can help solve those problems.

### Business Objective

After analyzing the data, we define our objectives as follows:

1. To predict if an employee will undergo attrition or not.
2. To identify factors which cause attrition.
3. Find unseen patterns in data that cause attrition.

### Stakeholders

The stakeholder in our Project is IBM (International Business machines corporation). IBM is a multinational technology company headquartered in New York.

### Business Benefit

Some of the benefits from this project are:

- Evaluation of employee needs, strengths and weaknesses.
- Based on employee profiling and company requirement the cost of hiring new employees can be reduced.
- Analysis of loss of important employees and their skillset.
- Measuring both financial and productive loss due to attrition.
- Prepare a contingency plan based on the prediction and insights provided by the learning model.

### Business Constraints

Constraints of this project are that the model we build will not be generalized. Since the dataset is obtained from Kaggle there are many factors which are missed out in the dataset and the other issue is factors affecting attrition change over time due to the addition of new employees. The model should be trained again and again to get accurate predictions.

## Analytical Understanding

To satisfy the business goals, we had to understand whether we could full fill the business goals using data mining and machine learning techniques by providing accurate results.

After proper analysis, we concluded that by using data mining and machine learning techniques we could achieve the task. Provided there is some hypothetical question that had to be answered.

- In order to predict attrition which machine learning model will be the best?
- How accurate the model would be?
- Would a non-technical stakeholder interpret the results obtained from the model?

# Data Understanding

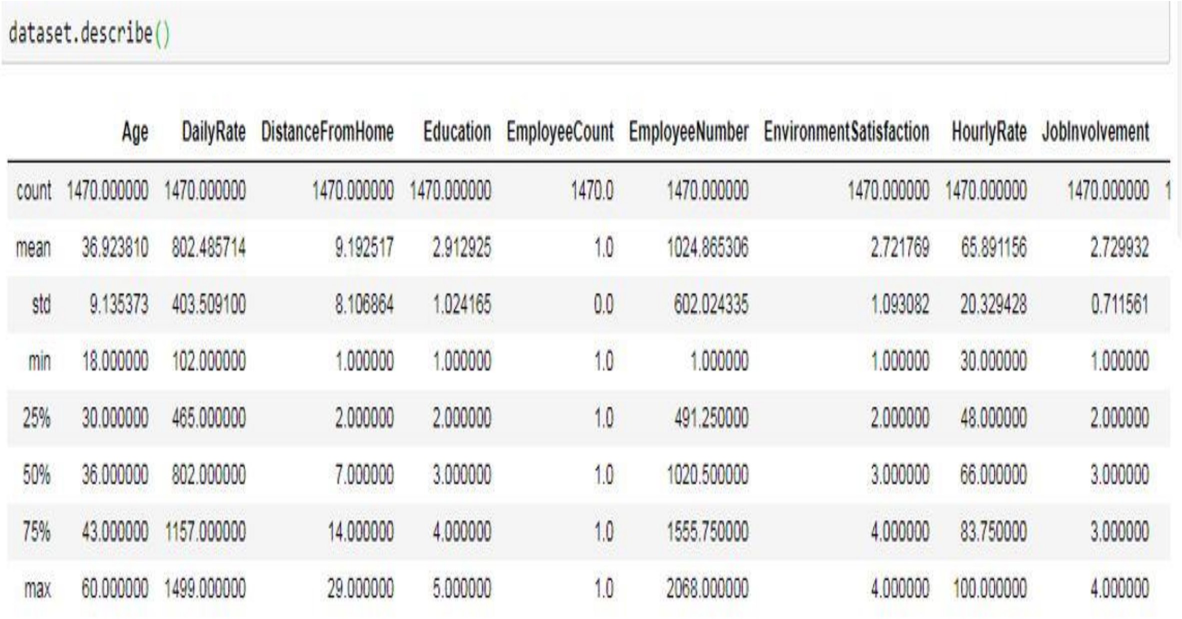
## Data Source

The dataset is taken from Kaggle. Which is an online open-source repository. The dataset link is provided below.

https://www.kaggle.com/ahmdel/ibm-hr-analytics-employee-attrition-performance

## Data Exploration

In order to get a better understanding of the data. We used python to do some descriptive analysis. The image below shows the description of each feature in the dataset.

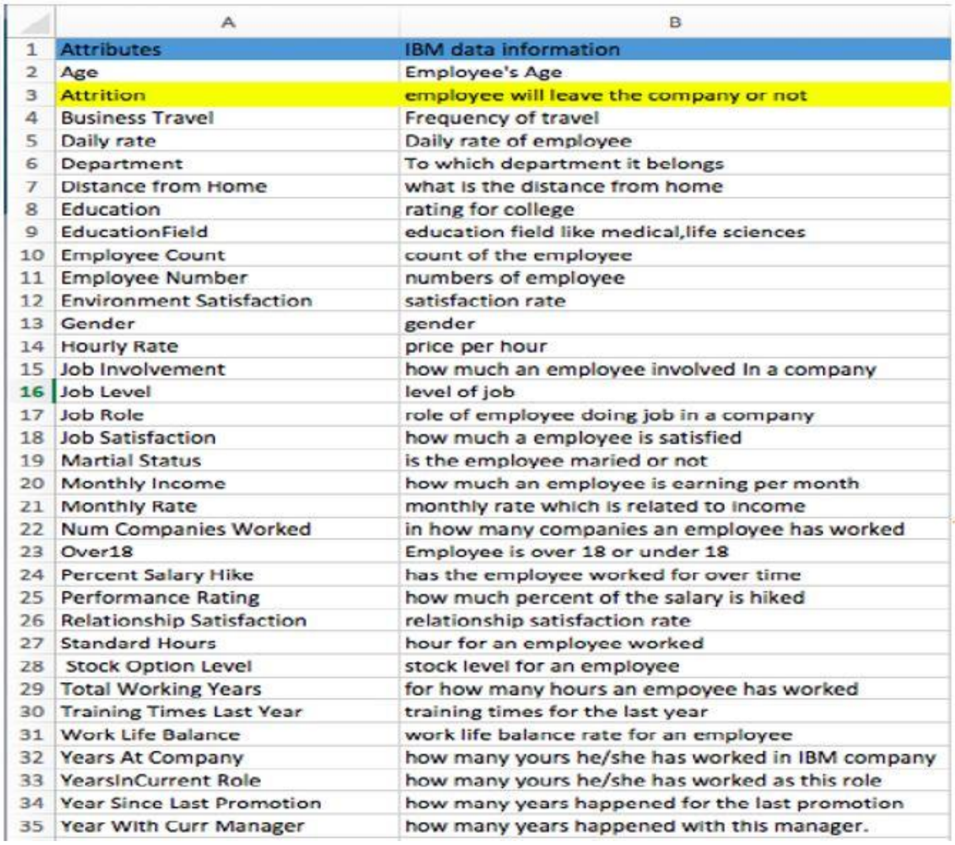


```
dataset.describe()
```

| | Age | DailyRate | DistanceFromHome | Education | EmployeeCount | EmployeeNumber | EnvironmentSatisfaction | HourlyRate | JobInvolvement |
|---|---|---|---|---|---|---|---|---|---|
| count | 1470.000000 | 1470.000000 | 1470.000000 | 1470.000000 | 1470.0 | 1470.000000 | 1470.000000 | 1470.000000 | 1470.000000 |
| mean | 36.923810 | 802.485714 | 9.192517 | 2.912925 | 1.0 | 1024.865306 | 2.721769 | 65.891156 | 2.729932 |
| std | 9.135373 | 403.509100 | 8.106864 | 1.024165 | 0.0 | 602.024335 | 1.093082 | 20.329428 | 0.711561 |
| min | 18.000000 | 102.000000 | 1.000000 | 1.000000 | 1.0 | 1.000000 | 1.000000 | 30.000000 | 1.000000 |
| 25% | 30.000000 | 465.000000 | 2.000000 | 2.000000 | 1.0 | 491.250000 | 2.000000 | 48.000000 | 2.000000 |
| 50% | 36.000000 | 802.000000 | 7.000000 | 3.000000 | 1.0 | 1020.500000 | 3.000000 | 66.000000 | 3.000000 |
| 75% | 43.000000 | 1157.000000 | 14.000000 | 4.000000 | 1.0 | 1555.750000 | 4.000000 | 83.750000 | 3.000000 |
| max | 60.000000 | 1499.000000 | 29.000000 | 5.000000 | 1.0 | 2068.000000 | 4.000000 | 100.000000 | 4.000000 |

The Dataset has 1470 records and 35 features. Each individual record holds information about an employee within the company IBM.

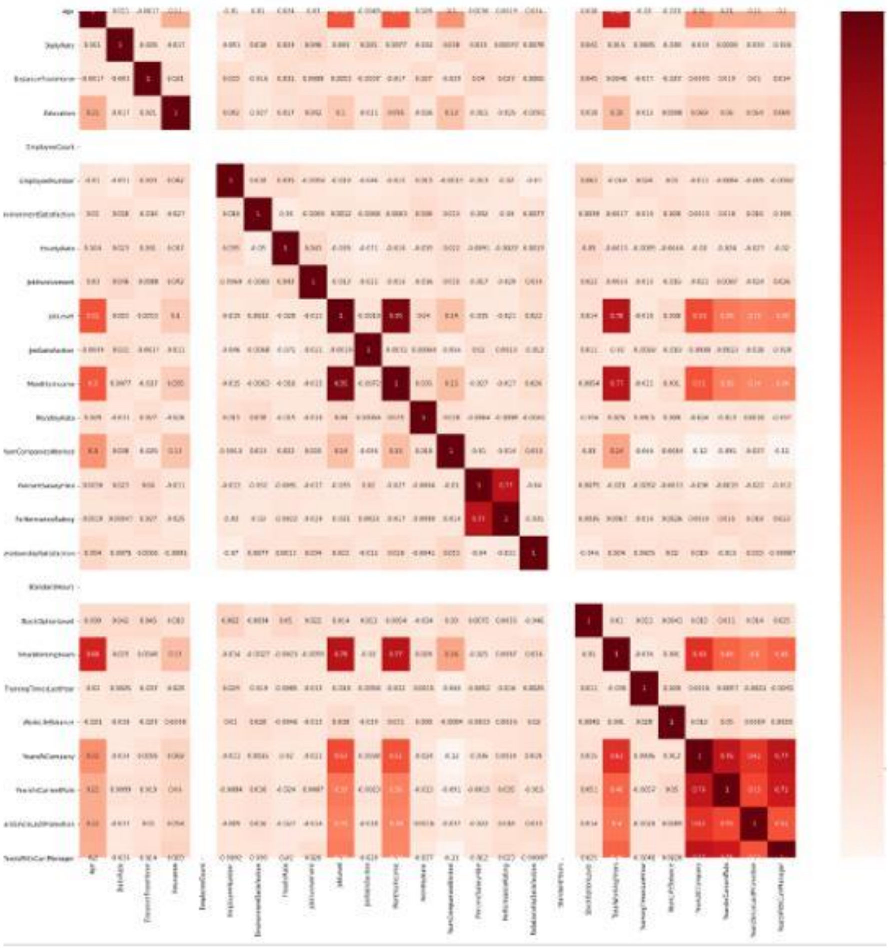
A brief description about each feature is shown below.



| | A | B |
|---|---|---|
| 1 | Attributes | IBM data information |
| 2 | Age | Employee's Age |
| 3 | Attrition | employee will leave the company or not |
| 4 | Business Travel | Frequency of travel |
| 5 | Daily rate | Daily rate of employee |
| 6 | Department | To which department it belongs |
| 7 | Distance from Home | what is the distance from home |
| 8 | Education | rating for college |
| 9 | EducationField | education field like medical,life sciences |
| 10 | Employee Count | count of the employee |
| 11 | Employee Number | numbers of employee |
| 12 | Environment Satisfaction | satisfaction rate |
| 13 | Gender | gender |
| 14 | Hourly Rate | price per hour |
| 15 | Job Involvement | how much an employee involved In a company |
| 16 | Job Level | level of job |
| 17 | Job Role | role of employee doing job in a company |
| 18 | Job Satisfaction | how much a employee is satisfied |
| 19 | Martial Status | is the employee maried or not |
| 20 | Monthly Income | how much an employee is earning per month |
| 21 | Monthly Rate | monthly rate which is related to income |
| 22 | Num Companies Worked | in how many companies an employee has worked |
| 23 | Over18 | Employee is over 18 or under 18 |
| 24 | Percent Salary Hike | has the employee worked for over time |
| 25 | Performance Rating | how much percent of the salary is hiked |
| 26 | Relationship Satisfaction | relationship satisfaction rate |
| 27 | Standard Hours | hour for an employee worked |
| 28 | Stock Option Level | stock level for an employee |
| 29 | Total Working Years | for how many hours an empoyee has worked |
| 30 | Training Times Last Year | training times for the last year |
| 31 | Work Life Balance | work life balance rate for an employee |
| 32 | Years At Company | how many yours he/she has worked in IBM company |
| 33 | YearsInCurrent Role | how many yours he/she has worked as this role |
| 34 | Year Since Last Promotion | how many years happened for the last promotion |
| 35 | Year With Curr Manager | how many years happened with this manager. |

There are 35 features in the dataset, we have chosen attrition as the output variable. Rest of the features will go through a feature selection process which will be explained in the later section.

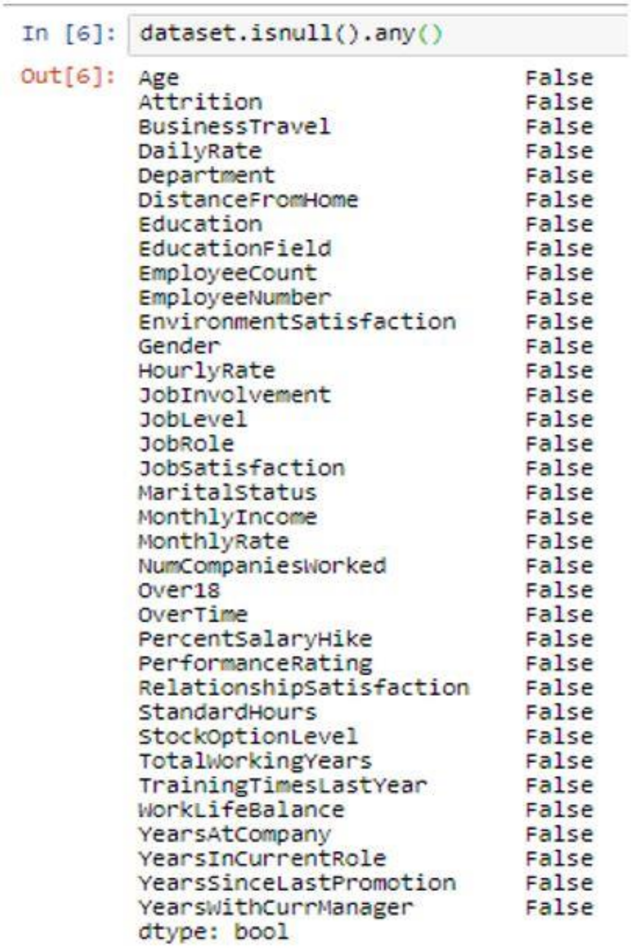
We plotted the correlation heat map using python to check which input features are correlated to each other and which input features have a high correlation to the output variable.



The input features which are highly correlated to each other may cause redundant information. Features which have a high positive (more than 0.7) or negative correlation (Less than –0.7) should be analyzed and reduced.

## Data Cleaning

Data cleaning is an important part before modelling. Missing or Null values can impact the performance of the algorithm. In order to check if there are any missing values, we used the python code. The below image shows the result obtained after executing the code.



```
In [6]: dataset.isnull().any()

Out[6]: Age                         False
        Attrition                   False
        BusinessTravel              False
        DailyRate                   False
        Department                  False
        DistanceFromHome            False
        Education                   False
        EducationField              False
        EmployeeCount               False
        EmployeeNumber              False
        EnvironmentSatisfaction     False
        Gender                      False
        HourlyRate                  False
        JobInvolvement              False
        JobLevel                    False
        JobRole                     False
        JobSatisfaction             False
        MaritalStatus               False
        MonthlyIncome               False
        MonthlyRate                 False
        NumCompaniesWorked          False
        Over18                      False
        OverTime                    False
        PercentSalaryHike           False
        PerformanceRating           False
        RelationshipSatisfaction    False
        StandardHours               False
        StockOptionLevel            False
        TotalWorkingYears           False
        TrainingTimesLastYear       False
        WorkLifeBalance             False
        YearsAtCompany              False
        YearsInCurrentRole          False
        YearsSinceLastPromotion     False
        YearsWithCurrManager        False
        dtype: bool
```

The above result depicted that there were no missing or null values with our dataset.

# Methodology

As mentioned in the introduction we are following the CRISP-DM methodology and we would implement this with the help of different machine learning and data mining techniques which are available in the rapid miner software.

## Software/Tools

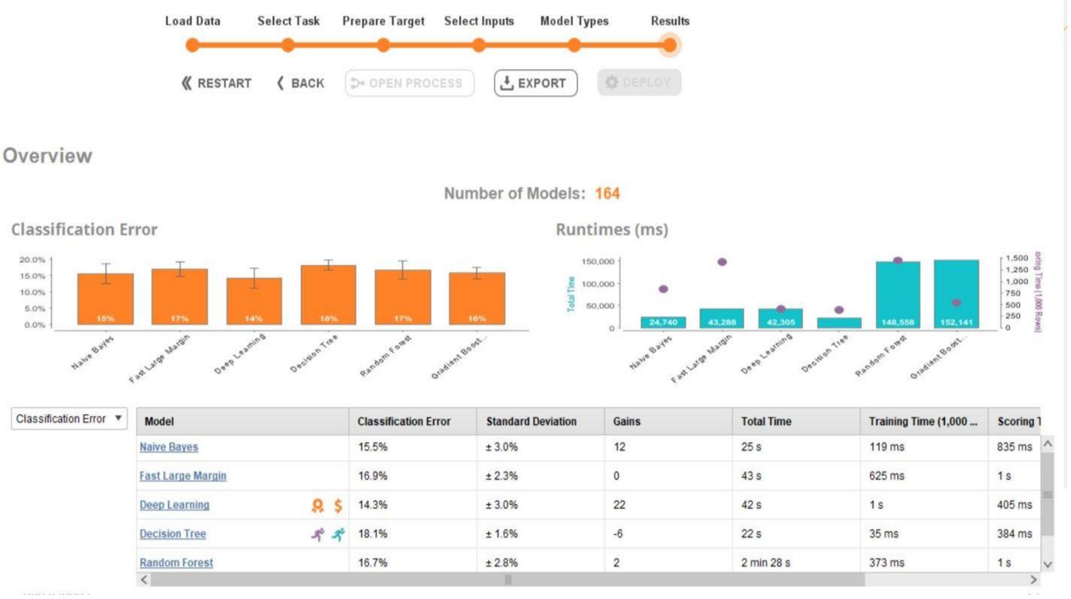
The Software or tools used to implement the process are listed below.

- Rapid Miner Studio.
- Python Jupyter

We used python to perform data exploration with some visualization and rapid miner for Data Cleaning, Data pre-processing, Feature engineering, Modelling and evaluation. Rapid miner provided a quick and efficient resource to implement the process.

## Testing and Algorithm Approach

We used auto model within rapid miner to select the top models which could be used in modelling. When we ran the auto model deep Learning was found to be the champion model whereas Decision tree was found to be the runner model. The below diagram shows the details about that.



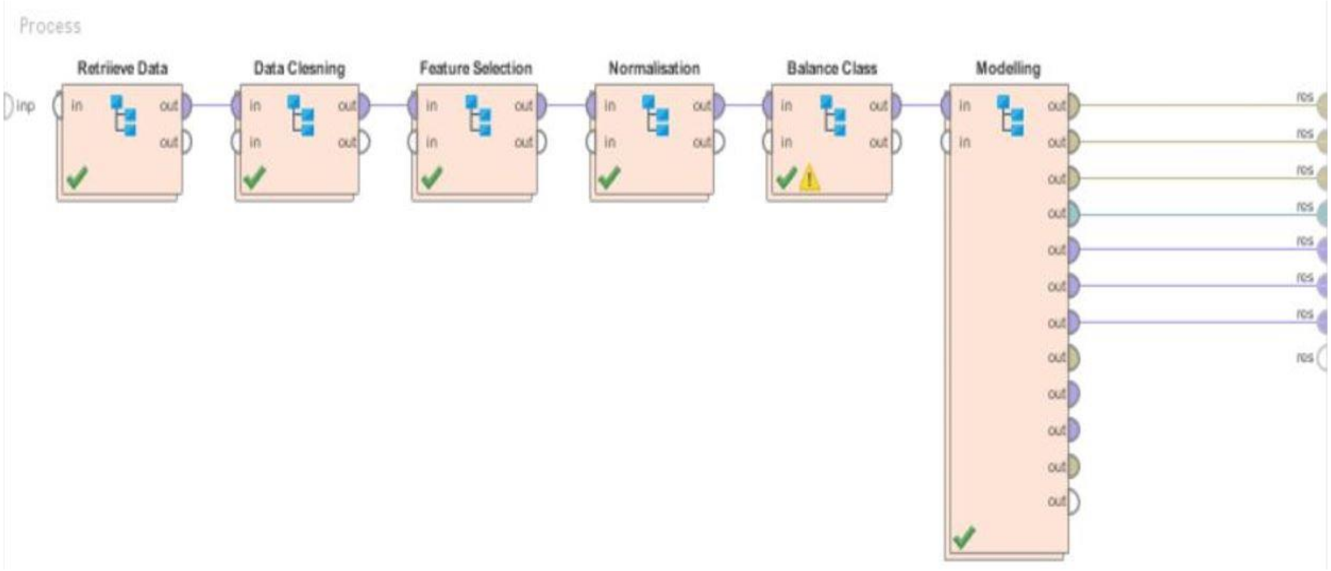
Number of Models: 164

| Model | Classification Error | Standard Deviation | Gains | Total Time | Training Time (1,000 ... | Scoring T |
|---|---|---|---|---|---|---|
| Naive Bayes | 15.5% | ± 3.0% | 12 | 25 s | 119 ms | 835 ms |
| Fast Large Margin | 16.9% | ± 2.3% | 0 | 43 s | 625 ms | 1 s |
| Deep Learning | 14.3% | ± 3.0% | 22 | 42 s | 1 s | 405 ms |
| Decision Tree | 18.1% | ± 1.6% | -6 | 22 s | 35 ms | 384 ms |
| Random Forest | 16.7% | ± 2.8% | 2 | 2 min 28 s | 373 ms | 1 s |

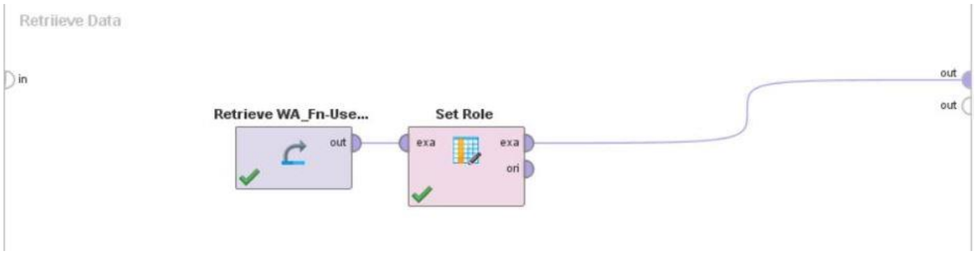
As we can see in the above diagram the ratio of classification error and runtime is much better in both deep learning and decision tree.

# Model Building and evaluation

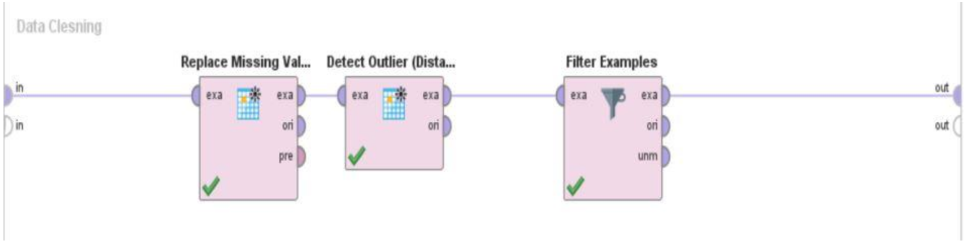
## Model Built



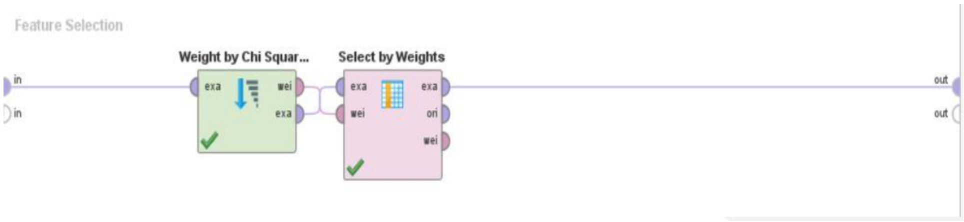
- Retrieve Data- This tool in rapid miner facilitates in reading the dataset from the device. We have used Retrieve operator to retrieve the CSV file and set role operator to initialize attrition as the output variable.



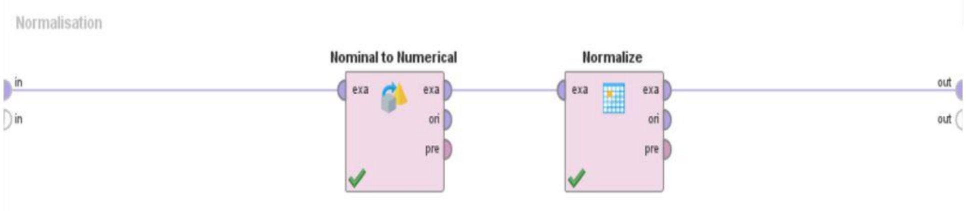
- Data Cleaning- In Data Cleaning we are detecting the top 100 outliers and removing them to reduce noise within the dataset. This will optimize the data and will provide better results from the models. Here we have used replace missing value operator to replace any missing value with the mean for numeric features and mode for categorical features.



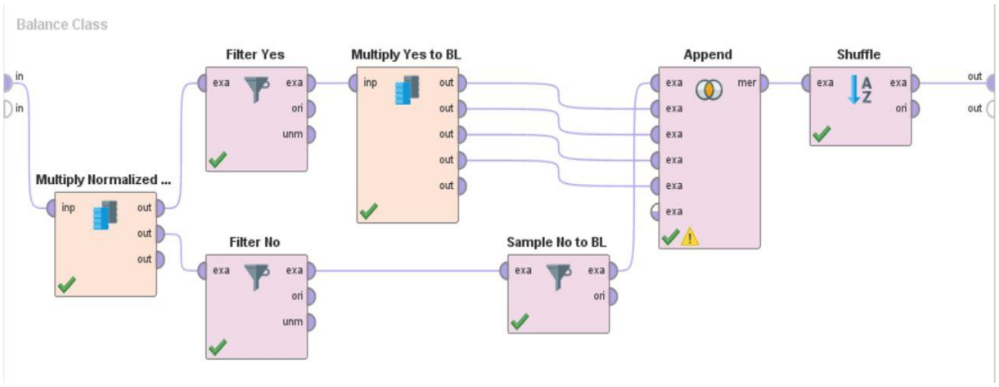
- Feature Selection-In Feature selection we are using chi-Square to filter out the top features that would improve the accuracy of the model. Weight by Chi-Square operator is used to score all the features and select by weights operator is used to filter out top 8 features.



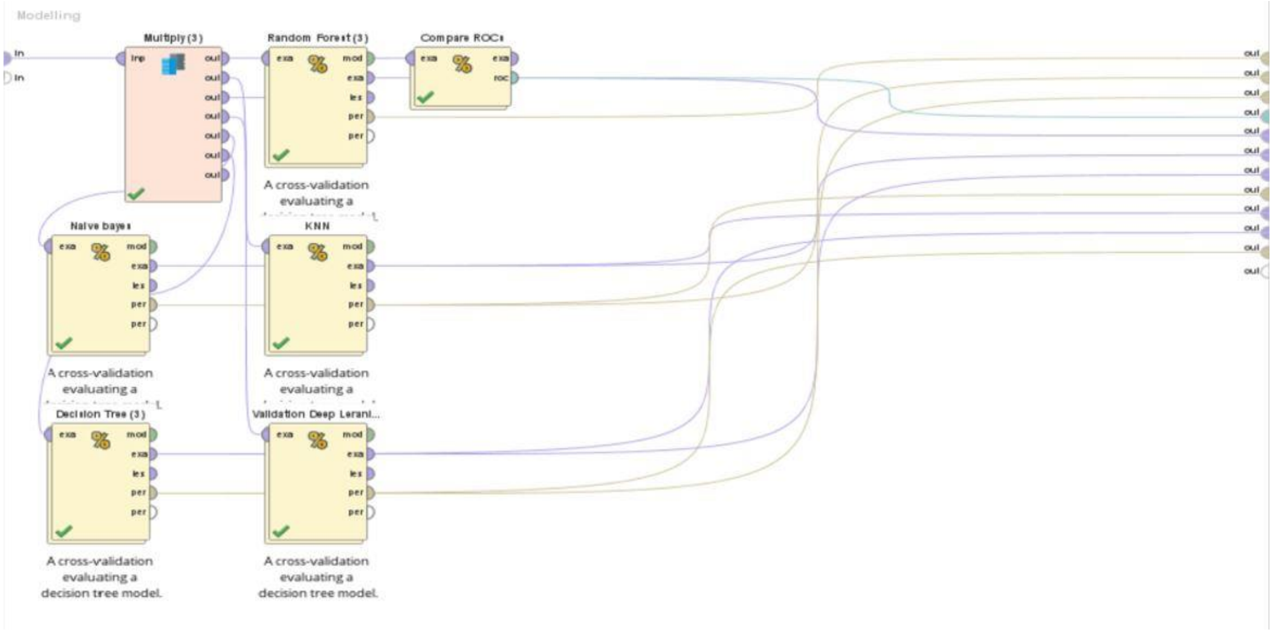
- Encoding and Normalization- We Have used Nominal to a numeric operator to convert all categorical columns to numbers. Normalize operator is used to bring all the features between the range of 0 to 1.



- Balance Class-Before feeding the data to the models we had to balance the output class. Imbalanced output classes will produce biased prediction towards the majority class. To balance we used a upsampling technique. Here we are using filter operator to first filter out Yes and No class in attrition then use multiply operator to increase the minority class Yes to the same level as No class. Append Operator is used to combine both records and shuffle operator is used to mix up the records.

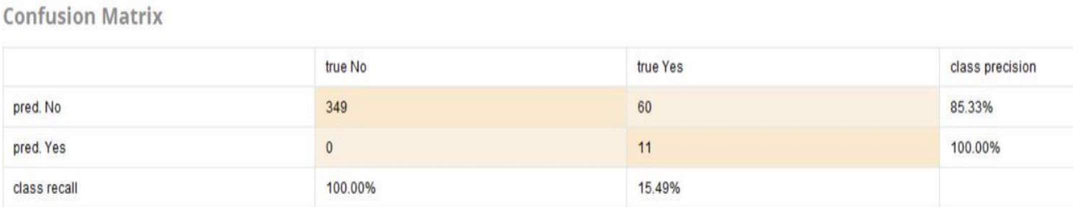


- Modelling- As we had seen in the auto model Deep Learning and Decision Tree showed great promise. But we also included Random Forest, Naïve Bayes and KNN to compare and check which algorithm would yield higher performance.


Modelling
in
in
Multiply (3)
inp
out
Random Forest (3)
Compare ROCs
exa
roc
mod
exa
tes
per
per
A cross-validation evaluating a
Naive bayes
KNN
Decision Tree (3)
Validation Deep Lerani...
A cross-validation evaluating a decision tree model.
A cross-validation evaluating a decision tree model.
out

## Evaluation

As we had seen earlier in the auto model Deep Learning had low error rate. The confusion matrix obtained from the auto model is shown below.

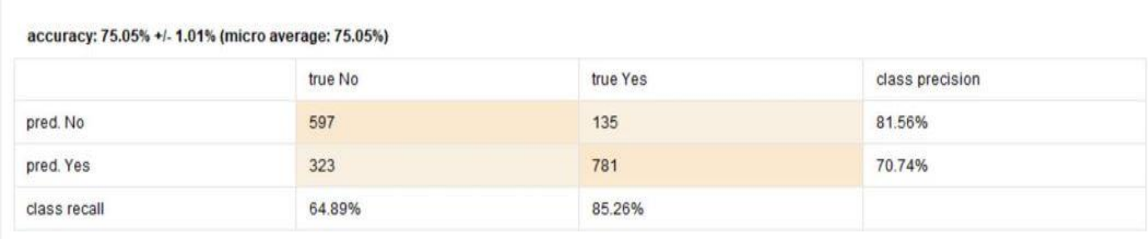


Confusion Matrix

| | true No | true Yes | class precision |
|---|---|---|---|
| pred. No | 349 | 60 | 85.33% |
| pred. Yes | 0 | 11 | 100.00% |
| class recall | 100.00% | 15.49% | |

The confusion matrix obtained from the auto model shows bad performance in Class Recall. Mainly because there is an imbalance in output class. But it had the best overall performance compared to other models.

During manual modelling after applying all the preprocessing techniques, feature engineering, class balance and normalization. we obtained the following confusion matrix for deep learning.

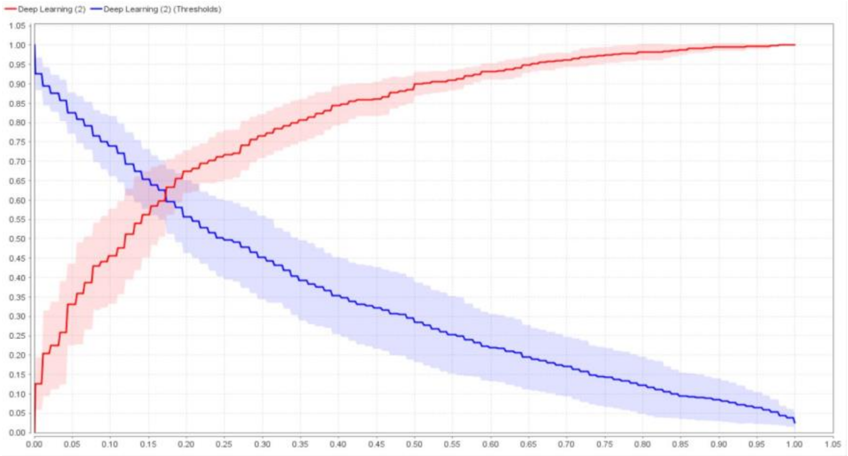


accuracy: 75.05% +/- 1.01% (micro average: 75.05%)

| | true No | true Yes | class precision |
|---|---|---|---|
| pred. No | 597 | 135 | 81.56% |
| pred. Yes | 323 | 781 | 70.74% |
| class recall | 64.89% | 85.26% | |

This confusion matrix shows better results compared to the auto model. The class recall values are balanced and precision values for No and Yes are 81.56% and 70.74% respectively, which is ideal. The accuracy is value is 75.05%. Overall, during manual modelling, we found that deep learning had the highest accuracy. So deep learning is the best model that can be chosen for satisfying the business objective.

We have plotted the roc graph to further visualize the performance. The diagram below the roc graph of Deep Learning.



As we can see from the diagram, we observe that the deep learning model is crossing the threshold and the Area under the curve is more than 0.75.

### Model Limitations

Even with good accuracy and low error rate, Deep Learning model has its disadvantages. Here are some of the disadvantages listed below-

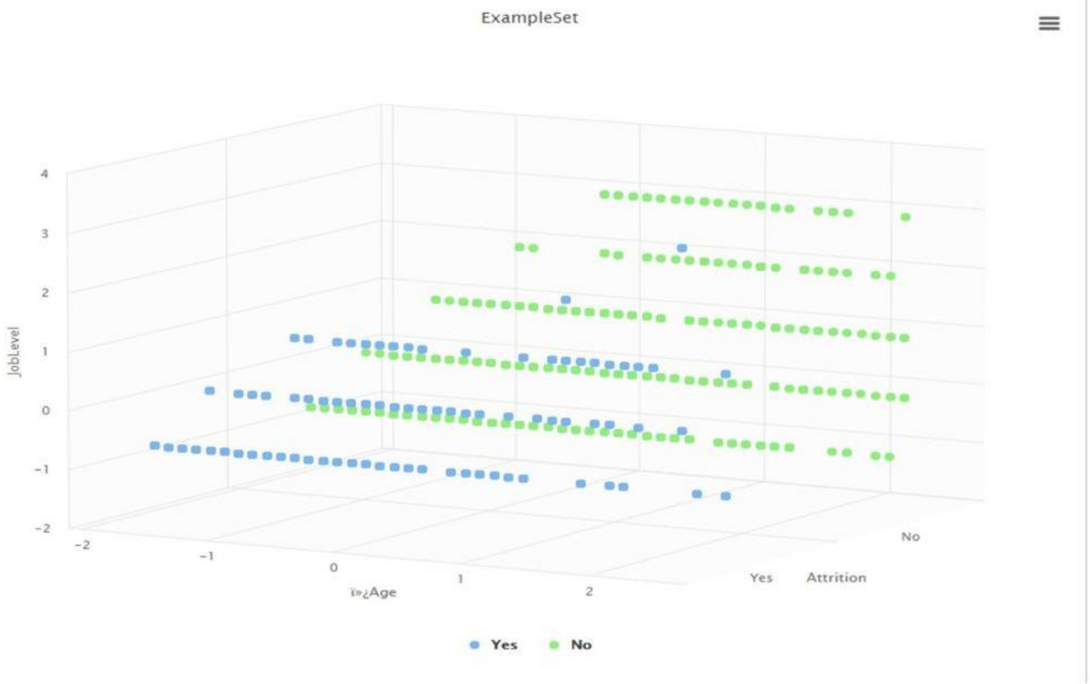
- Deep Learning is considered as a Blackbox, even though the results are good we don’t know how the allocation of weights is happening within the model.
- Deep Learning works best with big data since our dataset only contains 1470 records the analysis does provide a complete picture about the performance.
- Deep Learning models are computationally expensive.

## Deployment of Model

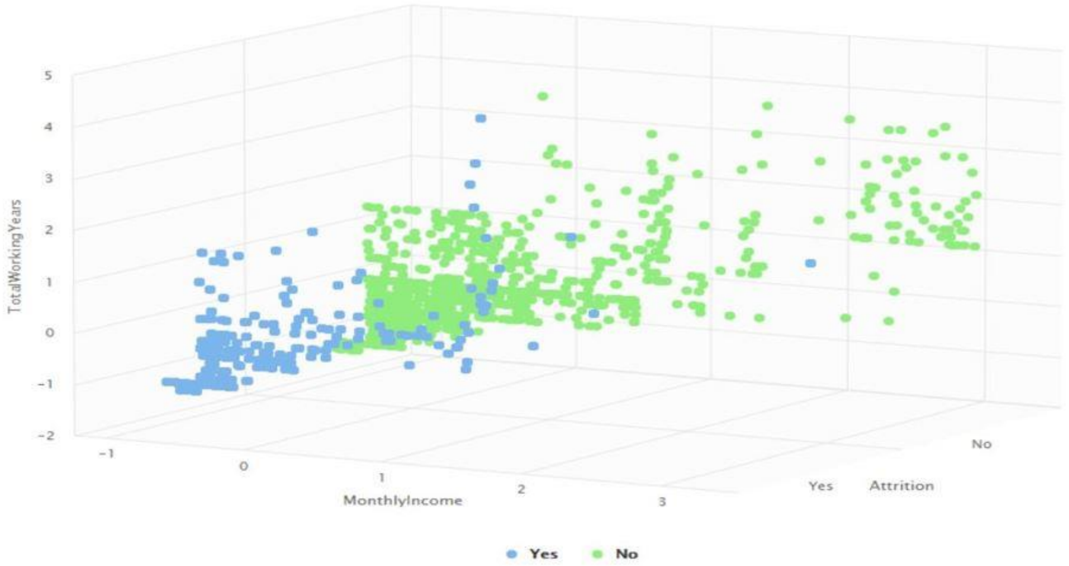
The final stage of our project deployment. There are a few steps which are necessary before deploying the process in a production environment.

### Business Validation

The goal of this project is to satisfy the business objective. To convince the non-technical stakeholder about the results obtained we have included some visualization.



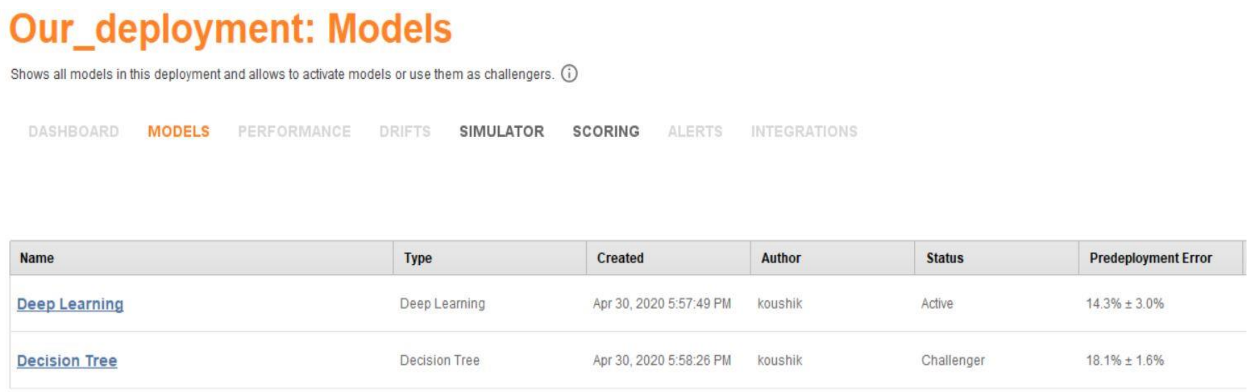
The above visualization shows that the attrition is high with younger employees who hold lower job levels. This suggests that people with less expertise and inexperience tend to leave the company. The company can take preventive measures to combat this problem.



The above visualization implies that employees with higher monthly income and employees with more experience tend to have a lower attrition rate. Whereas people with low income and less experience have higher attrition.

## Model Deployment

The Diagram below shows the deployment results we obtained in the rapid miner. This result is generated using the auto model feature within rapid miner.



Our_deployment: Models

Shows all models in this deployment and allows to activate models or use them as challengers. ⓘ

| Name | Type | Created | Author | Status | Predeployment Error |
|---|---|---|---|---|---|
| Deep Learning | Deep Learning | Apr 30, 2020 5:57:49 PM | koushik | Active | 14.3% ± 3.0% |
| Decision Tree | Decision Tree | Apr 30, 2020 5:58:26 PM | koushik | Challenger | 18.1% ± 1.6% |

The deployment feature within rapid miner can hold multiple models and can perform the same task with various models on the same dataset. It can be stored within the project repository and scores data.

Advantages of deployment:

- It keeps tract of essential data and model in a single place.
- Performance changes over time is detected and reported.
- It can be accessed by groups, working on the same project from different locations.
- Numerous Deployments can be held at the same Deployment location.

## Conclusion

We have performed the entire task of discovering the best model for our employee attrition dataset. Similarly, this approach can be implemented using rapid miner tool and python on a bigger industrial scale dataset. This approach can help different companies to save valuable time and prevent financial loss.

# References

Shearer, C. (2000). The crisp-dm model: the new blueprint for data mining. *journal of data warehouse*.