**Machine Learning Classification: HR Attrition Prediction**

Here are details of my next project on Machine Learning

Human Resources Analytics

Objective Find out why the best and most experienced employees leaving prematurely? Predict which valuable employees will leave next.  Here is the data: [HR\_comma\_sep2](https://rajivsworklife.files.wordpress.com/2017/09/hr_comma_sep2.xlsx)

**Exploratory Data Analytics**

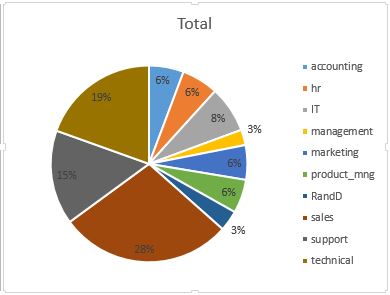
The first step was to check if there were any records with blank fields.  There were none.

The Label field is “Left” and it indicates those who left.  Let us start by exploring the correlation between the label field and the feature fields.

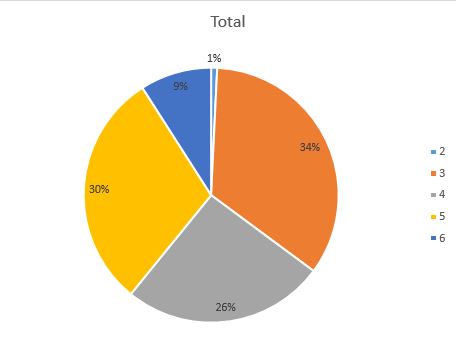
|  |  |
| --- | --- |
| Feature | Correlation |
| satisfaction\_level | -0.388374983 |
| last\_evaluation | 0.00656712 |
| number\_project | 0.023787185 |
| average\_montly\_hours | 0.071287179 |
| time\_spend\_company | 0.144822175 |
| Work\_accident | -0.154621634 |
| promotion\_last\_5years | -0.061788107 |
| salary | -0.157897791 |

There is a strong negative correlation between whether a person is leaving to his/her Satisfaction Level, Salary, If the person met with a Work Accident and if there has been a promotion in the last 5 years.

Also the department with the highest attrition is the Sales department with a figure of 28%, followed by the Technical department and Support department – at 20% and 16% respectively.



Also the longer a person stays with the company the less likely that he or she would leave the organization



As you can see a person with 3 years in the organization is the most vulnerable to leave, followed by 5 years and then by 4 years in the organization.  People with more than 6 years experience have not been found to leave the organization.

**Prediction Approach**

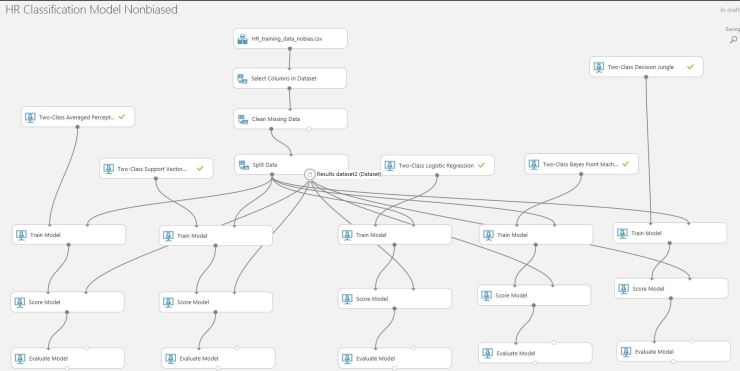
The objective is to predict the list of people who are most likely to leave the organization.  To train the model we split the data into the training set and the testing set.

The training set has data of everybody who has left and also an equal number of those who have not left.  To select those who have not left – we would choose those who are still within the organization – but who have the lowest satisfaction\_level, salary, Work\_accident and promotion\_last\_5years (in that order).  From that subset of data – we choose an equal number of records in **reverse order**.  The reason to choose the reverse order is so that we can create a training dataset that is not biased.  So the reversed order subset of data would contain those with the highest satisfaction level, salary, no work\_accidents and have been promoted in the last 5 years.

Another set of training data is also created with a combination of data of everybody who has left and a set of biased data from those who have not left (those who are quite likely to leave based on high correlation).

Both the above training data is used to create a model that is more highly accurate.

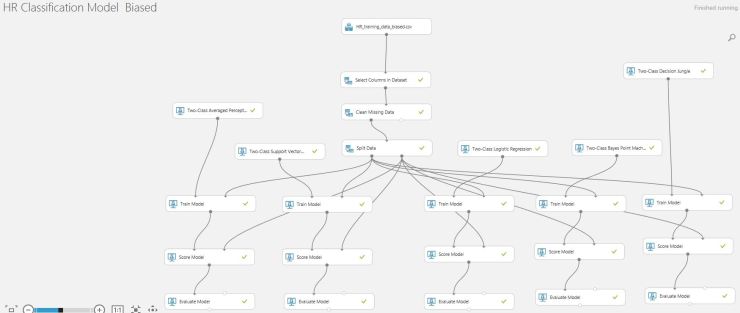
We use Microsoft Azure Machine Learning Studio to play with various algorithms to create the best one:



There are five algorithms chosen above.  The performance metrics for each of the above is given below:

|  |  |  |  |
| --- | --- | --- | --- |
| Name of Algorithm | AUC | Accuracy | Precision |
|  |  |  |  |
| Two-Class Averaged Perceptron | 0.749 | 0.682 | 0.683 |
| Two-Class Support Vector Machine | 0.673 | 0.599 | 0.589 |
| Two-Class Logistic Regression | 0.745 | 0.677 | 0.676 |
| Two-Class Bayes Point Machine | 0.744 | 0.677 | 0.679 |
| Two-Class Decision Jungle | 0.991 | 0.969 | 0.990 |

Next we change the input dataset to a biased dataset and then we gauge how the above five perform



|  |  |  |  |
| --- | --- | --- | --- |
| Name of Algorithm | AUC | Accuracy | Precision |
| Two-Class Averaged Perceptron | 0.853 | 0.800 | 0.745 |
| Two-Class Support Vector Machine | 0.835 | 0.778 | 0.725 |
| Two-Class Logistic Regression | 0.853 | 0.801 | 0.744 |
| Two-Class Bayes Point Machine | 0.851 | 0.792 | 0.731 |
| Two-Class Decision Jungle | 0.992 | 0.974 | 0.991 |

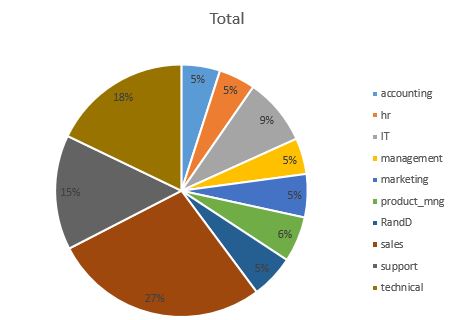
It therefore appears that the biased dataset seems to create a better model.  Of the above five, we can choose the top three in the below order:

First: Two-Class Decision Jungle

Second: Two-Class Logistic Regression

Third: Two-Class Averaged Perceptron

Decision Jungle Model was chosen as the best.  Upon analysis of the predicted outcome – the department wise pattern was very similar to modelling data:



27% from Sales, 18% from Technical and 15% from support.

(The model is [Decision Jungle Model Biased](https://rajivsworklife.files.wordpress.com/2017/09/decision-jungle-model-biased.xlsx) along with the solutions API key: CprmDLspC+PeGrwL3269PD+VnsaBR6lsS1Bgi9Dh+xt0LGvU1NRnYVoUDSj+GgwyBG7ezl8ermhf9gmyn0pDXg==).  The model predictions were pretty accurate when I ran it across the validation dataset.