Employee Attrition Data Analysis Using Keras and Bokeh

IBM HR Analytics Employee Attrition

Abstract

Employee Attrition is a major problem in most of the companies. Every company wants to know the factors that lead to Employee leaving a company. The following report focuses on highlighting some of the major factors responsible for employee attrition using Keras and Bokeh. Keras is used to create a model that can help in predicting which employees can leave the company in the future. Bokeh is used to plot graphs which provides various insights into the data. The dataset used for this research is IBM HR Analytics Employee Attrition. It can be downloaded from Kaggle (Link shown in Reference Section). The purpose of this report is to provide suggestions where a company can take action and reduce Employee attrition based on different department. Initially, a brief introduction of the dataset is provided then some of the functions of Bokeh and Keras that are used in the analysis are introduced. Finally, some insights into dataset are given using bokeh and the conclusion that we can derive from them.

IBM HR Analytics Employee Attrition

Before starting with the analysis, it is important to understand the structure of the dataset. The dataset consists of 35 columns which are described in the below table.

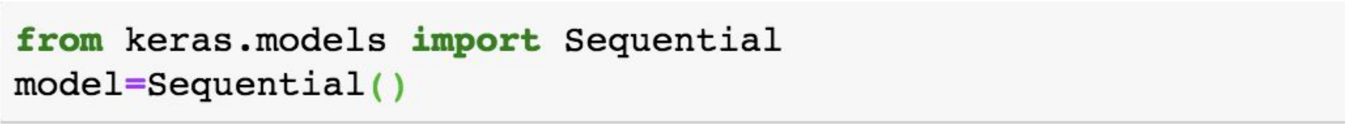
Parameter Name	Description	Parameter Name	Description
Age	Age of an Employee.	Attrition	A value indicating whether the given employee has left the company or not.
Business Travel	A parameter indicating the amount of travel done by the employee	Daily rate	The amount allocated per day to the employee for travel purpose.
Department	The department in which the employee is working.	Distance from home	The distance traveled by an employee from his home to reach the office.
Education	A numerical value indicating the level of education of an employee. 1 'Below College' 2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor'	Education Field	The field in which the employee has taken his education.
Employee Count	Count of an employee.	Employee Number	Unique Employee Id.
Environment Satisfaction	A numerical value indicating employee satisfaction. 1 'Low' 2 'Medium' 3 'High' 4 'Very High'	Gender	Gender of the Employee.

Parameter Name	Description	Parameter Name	Description
Hourly Rate	Salary given to an employee on an Hourly basis	Job involvement	A numerical value indicating how demanding is the job.1 'Low' 2 'Medium' 3 'High' 4 'Very High'
Job Level	A numerical value indicating the job level of an employee.	Job Role	The role of the employee in the company.
Job satisfaction	A numerical value indicating job satisfaction of an employee. 1 'Low' 2 'Medium' 3 'High' 4 'Very High'	Marital Status	The marital status of an employee.
Monthly Income	The monthly salary earned by the employee.	Monthly Rate	The monthly rate of an employee including his salary and other expenditure.
NumCompaniesWorked	The number of companies in which the employee has worked	Over 18	A value indicating whether the employee age is above 18 or not.
OverTime	A value indicating whether the employee has done overtime or not.	Percent Salary Hike	The percentage of salary hike given to an employee
Performance Rating	A rating given to an employee based on his performance. 1 'Low' 2 'Good' 3 'Excellent' 4 'Outstanding'	Relationship satisfaction	A numerical value indicating the relationship satisfaction of an employee with his Manager.1 'Low' 2 'Medium' 3 'High' 4 'Very High'
Standard hours	Standard Working Hours of the department in which the employee is working.	Stock Option Level	The number of stocks given to the employee.
Total Working Years	Total Experience of an employee	Training Times Last Year	Amount of Training taken by employee last year
WorkLifeBalance	A numerical rating indicating the work-life balance of an employee. 1 'Bad' 2 'Good' 3 'Better' 4 'Best'	YearsAtCompany	The number of years the Employee has worked in the company.
YearsInCurrentRole	Number of years the employee has worked in the current role	Years Since Last Promotion	Number of years the employee has not received the promotion.

YearsWithCurrManager	The number of years the employee has worked with the current manager.		

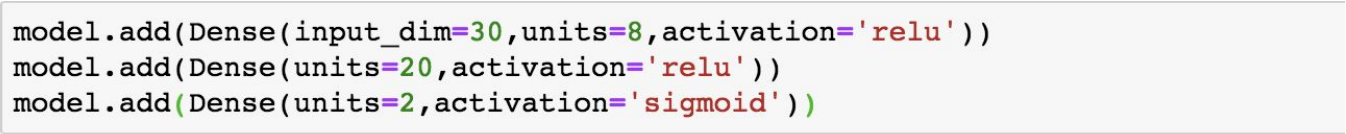
Keras

Keras is a high-level neural networks API, written in Python and capable of running on top of TensorFlow, CNTK, or Theano. It was developed with a focus on enabling fast experimentation. Being able to go from idea to result with the least possible delay is key to doing good research. The core data structure of Keras is a model, a way to organize layers. The simplest type of model is the Sequential model, a linear stack of layers. The following snapshot shows how to create a Sequential model.



```
from keras.models import Sequential
model=Sequential()
```

The layers will be stacked using .add function. It is shown in the following snapshot:-

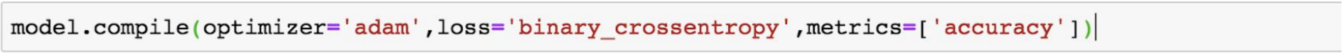


```
model.add(Dense(input_dim=30,units=8,activation='relu'))
model.add(Dense(units=20,activation='relu'))
model.add(Dense(units=2,activation='sigmoid'))
```

For employee attrition dataset, three layers are created. The first layer is the input layer where 30 independent variables are provided as input. It is important to note that the number of the parameter in our dataset is 35 but only 30 variables are provided as input to the neural network. The attrition value is the dependent variable so it is not considered while training the model. The other independent variable like Employee Count, Employee Number, Over 18 and standard hours are not considered while training the model since they do not have influence in predicting the final output.

The number of nodes in the corresponding hidden layers are chosen empirically. The last layer will have two nodes indicating whether the employee will leave the company or not. The second parameter indicates the activation function that will be used in each layer. For our dataset, **RELU** is used in the first two layers while in the output layer the **sigmoid** function is used. Since neural networks are sensitive to changes, the activation function is used which helps in smoothing the output i.e. small changes in input does not lead to large changes in output.

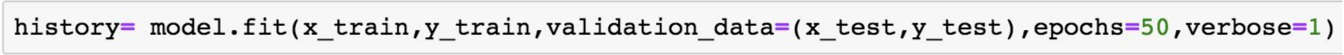
After initializing the model, the model is compiled which is shown in the following snapshot:-



```
model.compile(optimizer='adam',loss='binary_crossentropy',metrics=['accuracy'])
```

The optimizer defines how the weights will be updated in training data. For our dataset, **Adam** is used as an optimization function which is the variant of classical stochastic gradient descent and combines the advantage of both Adaptive Gradient Algorithm (AdaGrad) and Root Mean Square Propagation (RMSProp).
The Loss function is an important part of deep learning, which is used to measure the inconsistency between the predicted value and actual label. In our case, the binary_crossentropy is used and the metric used for model evaluation is accuracy.

The training is started using a .fit function which is shown in the following snapshot:-



```
history= model.fit(x_train,y_train,validation_data=(x_test,y_test),epochs=50,verbose=1)
```

one epoch = one forward pass and one backward pass of all the training examples.

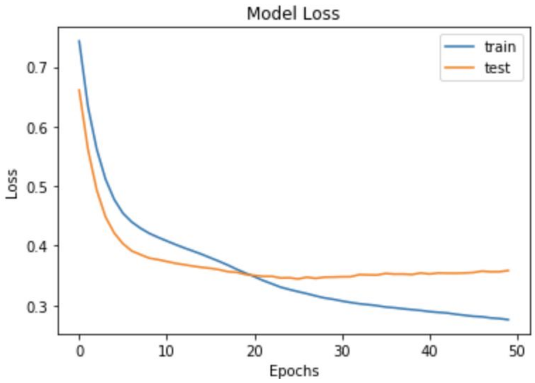


Fig 1: Graph showing model loss on the training and test set for each epoch.

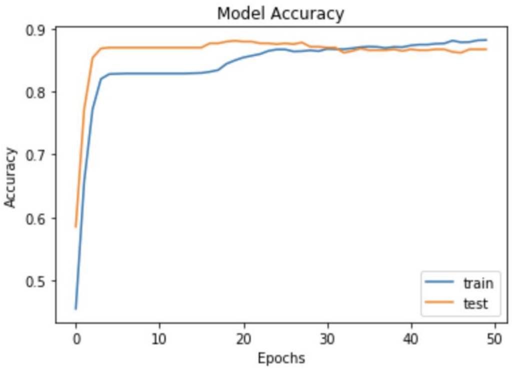
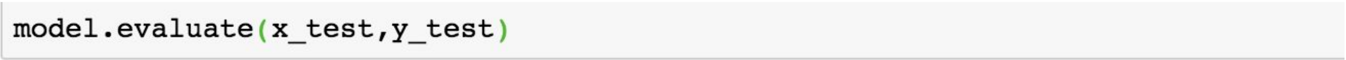


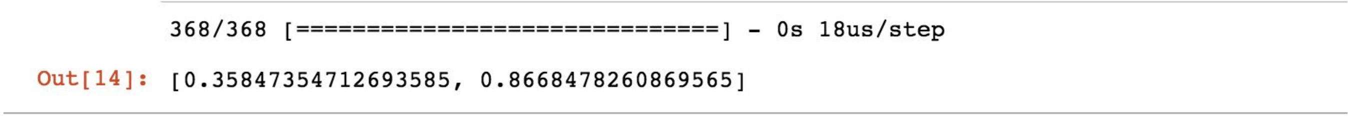
Fig 1: Graph showing model Accuracy on the training and test set for each epoch.

The model evaluation is done on the test set as shown in the following snapshot:-



```
model.evaluate(x_test,y_test)
```

The final accuracy obtained while evaluating on test set was 86% as shown in the following snapshot.



```
         368/368 [==============================] - 0s 18us/step
Out[14]: [0.35847354712693585, 0.8668478260869565]
```

Visualization using Bokeh

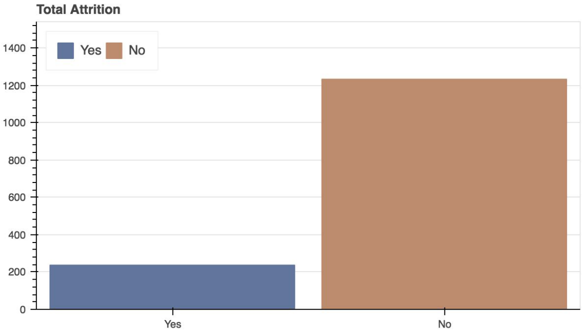


Fig 1: Total Attrition of employee

Analysis

The bar chart shows the total attrition of employees. The number of employees who have left the company is quite low (237) as compared to employees who are working in the company(1233).

Yes, Percentage = 1233/(237 + 1233) = 83 %

No, Percentage = 237/(237 + 1233) = 17 %

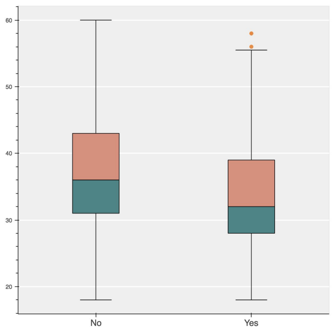


Fig 2: Attrition of employee based on Age

Analysis

The box chart shows the minimum, median and maximum ages of employees who have left the company and those who are currently working in the company

It is important to note that the median as well the maximum age of the people with 'No' attrition is higher than that of the 'Yes' category. This shows that people with higher age have lesser tendency to leave the organization which makes sense as they may have settled in the organization.

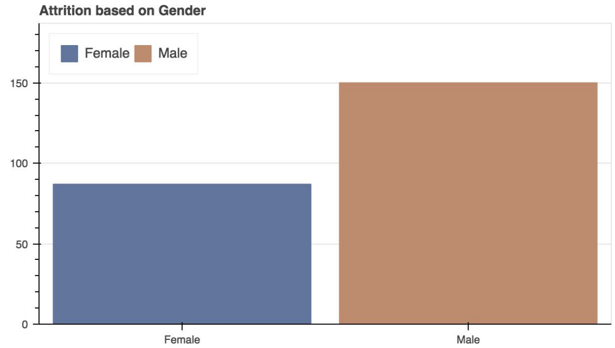


Fig 3: Attrition of employee based on Gender

Analysis

The bar chart shows attrition based on gender. It can be observed that attrition of males is almost double as compared to number of Females.

Male Attrition = 150/237 = 63%

Female Attrition = 87/237 = 37%

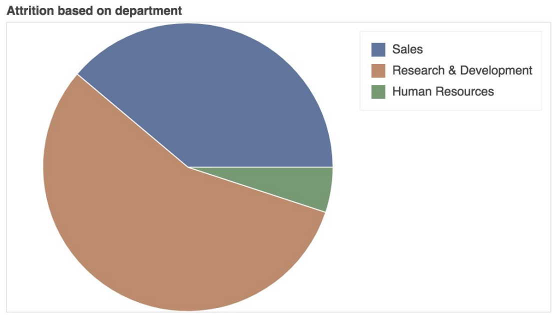


Fig 4: Attrition of employee in different departments

Analysis

The pie chart shows the attrition of employees based on department. It can be observed that research and Development (133) has the highest number of attrition as compared to Sales (92) and Human Resources (12).

Sales Attrition = 92 / 237 = 38%
Research & Development Attrition = 133/ 237 = 56%
Human Resource Attrition = 12 / 237 = 5%

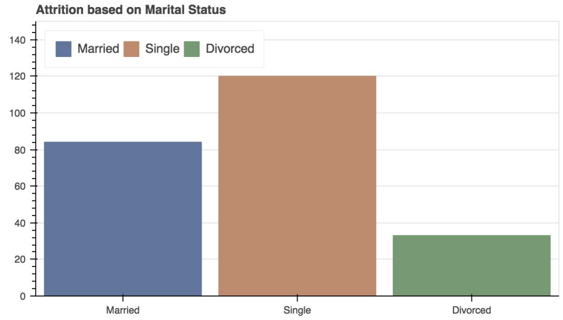


Fig 5: Attrition of employee based on Marital Status

Analysis

The bar chart shows the attrition of employees based on marital Status. It can be observed from the graph the number of attrition For Single is more as compared to Married and Divorced.

Married Attrition = 84/237 = 35%
Single Attrition = 120/237 = 51%
Divorced Attrition = 33/237 = 14%

Now we will find reasons for attrition in Research and Development based on different parameters like business travel, Ove
Job Involvement,

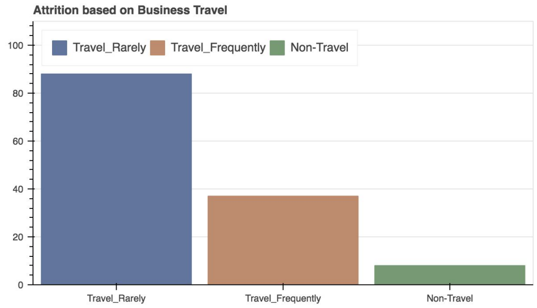


Fig 5: Attrition of employee based on Business Travel in Research and development department

Analysis

The bar chart shows Attrition of employee based on Business Travel in Research and development department.It can be observed that employees who travel rarely (88) has more amount of attrition as compared to who travel frequently(37) and non-travelers (8).

Travel Rarely = 88/133 = 66%
Travel Frequently = 37/133 = 28 %
Non-Travel = 8/133 = 6%

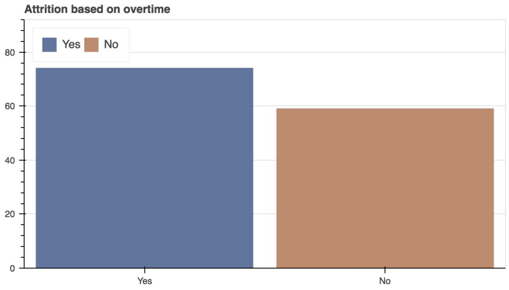


Fig 7: Attrition of employee based on overtime.

Analysis

The bar chart shows Attrition of employee based on Overtime in Research and development department. It can be observed that employees who are doing more overtime(74) have more attrition rate compared to those who leave on time(59).

Attrition for Overtime = 74/133 = 56%
Attrition for who are not doing Overtime = 59/133 = 44%

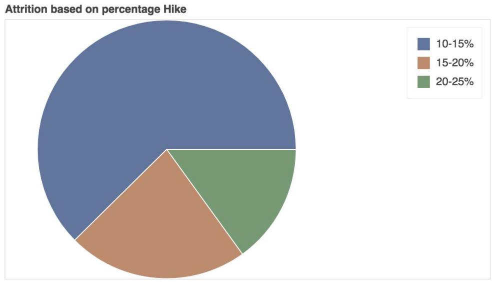


Fig 8: Attrition of employee based on Salary Hike

Analysis

The pie chart shows attrition of employees based on percentage salary hike in research and development department. It can be observed that the attrition rate was high for employees who got lower percentage hike.
Attrition in percentage for different percentage hike are as follows:-

10-15% Hike = 83/133 = 62%, 15-20% Hike = 30/133 = 23%
20-25% Hike = 20/133 = 15%

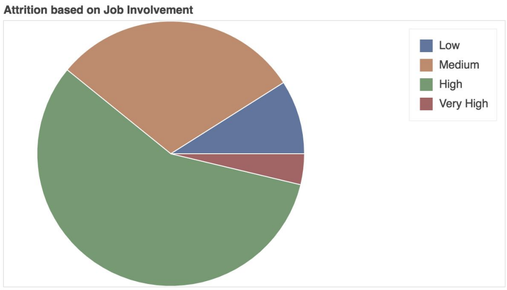


Fig 7: Attrition of employee based on Job Involvement

Analysis

The pie chart shows Attrition based on Job Involvement in Research and development department. It can be observed that the attrition rate was high for employees having more job involvement.

Attrition in percentage for different Job Involvement are as follows:-

Low = 12/133 = 9% , Medium = 40 /133 = 30%, High = 57%
Very High = 3%

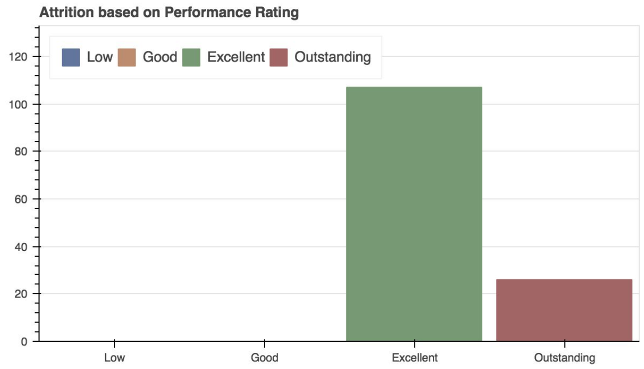


Fig 7: Attrition of employee based on Performance rating.

Analysis

The bar chart shows attrition based on performance rating for research and development department. It can be observed that the High performing employees were the most to leave the company.

Attrition in percentage for different Performance rating are as follows:-

Low = 0%, Good =0%, Excellent = 80%, Outstanding = 20%

Major reasons for attrition in entire company

- The environment satisfaction was found to be one of the major reasons for employee attrition. The company needs to change some policies to improve the work environment.
- The average age of people leaving the company were young, the company needs to devise some policies that can help in reducing the number of attrition. (**Refer** Fig 2)
- The major concern found in all the department was the employees who had good performance were leaving the company. The company needs to devise a policy, that they can help in retaining high performers.

Major reasons for attrition department wise

Different departments had different reasons for attrition. Some of the major reasons for attrition in each department and the suggestion for improvisation is shown in the following table.

Department Name	Major Reasons for Attrition	Suggestions
Research and development (56%)	Percentage Salary Hike (62%)	The company needs to increase the salary percentage bracket for average performers.
Sales (38%)	Business Travel (73%)	Employees who travelled rarely had major amount of attrition. The company needs to provide opportunity for the sales people to travel more and explore business opportunities.
Human Resource Attrition (5%)	Percentage Salary Hike (83%)	The company needs to increase the salary percentage bracket for average performers.

References

https://keras.io/
https://bokeh.pydata.org/en/latest/
https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset/kernels