

# **Corporate Employee Attrition Analysis**

## **A PROJECT COMPONENT REPORT**

*Submitted by*

**TEAM ID : PNT2022TMID15667**

## TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
1	ACKNOWLEDGEMENTS	3
2	OBJECTIVE	4
3	DESCRIPTION OF PROJECT	5
4	METHODOLOGY	5
5	VISUALIZATIONS AND MODEL EVALUATION	8
6	REFLECTION ON THE PROJECT	13
7	CONCLUSION	14
8	LINK TO CODE AND EXECUTABLE FILES	15

## ACKNOWLEDGEMENTS

The Nalaiya Thiran opportunity I had with IBM in the Data Analytics domain was a great chance for learning and professional development. Therefore, I am also grateful to all the professionals who led me through this internship period.

I express my deepest thanks to my industry mentor for taking part in useful decision & giving necessary advice and guidance and arranged all facilities to make internship easier. I choose this moment to acknowledge his contribution gratefully.

It is my radiant sentiment to place on record my best regards, deepest sense of gratitude to our college faculty mentor for his careful and precious guidance which were extremely valuable for my study both theoretically and practically.

I perceive as this opportunity as a big milestone in my career development. I will strive to use gained skills and knowledge in the best possible way, and I will continue to work on their improvement, in order to attain desired career objectives.

Sincerely,

**TEAM ID : PNT2022TMID15667**

# OBJECTIVE

- The objective of this project is to predict the attrition rate for each employee, to find who's more likely to leave the organization.
- It will help organization to find ways to prevent attrition or plan the hiring of the new candidate.
- Attrition proves to be a costly and time-consuming problem for the organization and it also leads to the loss of probability.
- The scope of the project extends to companies to all industries

# DESCRIPTION OF PROJECT

In this Project, it is required to clean and sanitize the dataset. Then, train the dataset to predict the attrition rate of the employees in an organization

## METHODOLOGY

### **1. Business Understanding:**

Before solving the problem in the Business domain, it needs to be understood properly. Business understanding forms a concrete base, which further leads to easy resolution of queries. We should have the clarity of what is the exact problem we are going to solve.

### **2. Analytic Understanding:**

Based on the above business understanding one should decide the analytical approach to follow. The approaches can be of 4 types: Descriptive approach (status and information provided), Diagnostic approach (A.K.A statistical analysis, what is happening and why it is happening), Predictive approach (it forecasts on the trends or future events probability) and Prescriptive approach (how the problem should be solved actually).

### **3. Data Requirements:**

The above chosen analytical method indicates the necessary data content, formats, and sources to be gathered. During the process of data requirements, one should find the answers for questions like 'what', 'where', 'when', 'why', 'how' & 'who'.

#### **4.Data Collection:**

Data collected can be obtained in any random format. So, according to the approach chosen and the output to be obtained, the data collected should be validated. Thus, if required one can gather more data or discard the irrelevant data.

#### **5.Data Understanding:**

Data understanding answers the question “Is the data collected representative of the problem to be solved?”. Descriptive statistics calculates the measures applied over data to access the content and quality of matter. This step may lead to reverting the back to the previous step for correction.

#### **6.Data Preparation:**

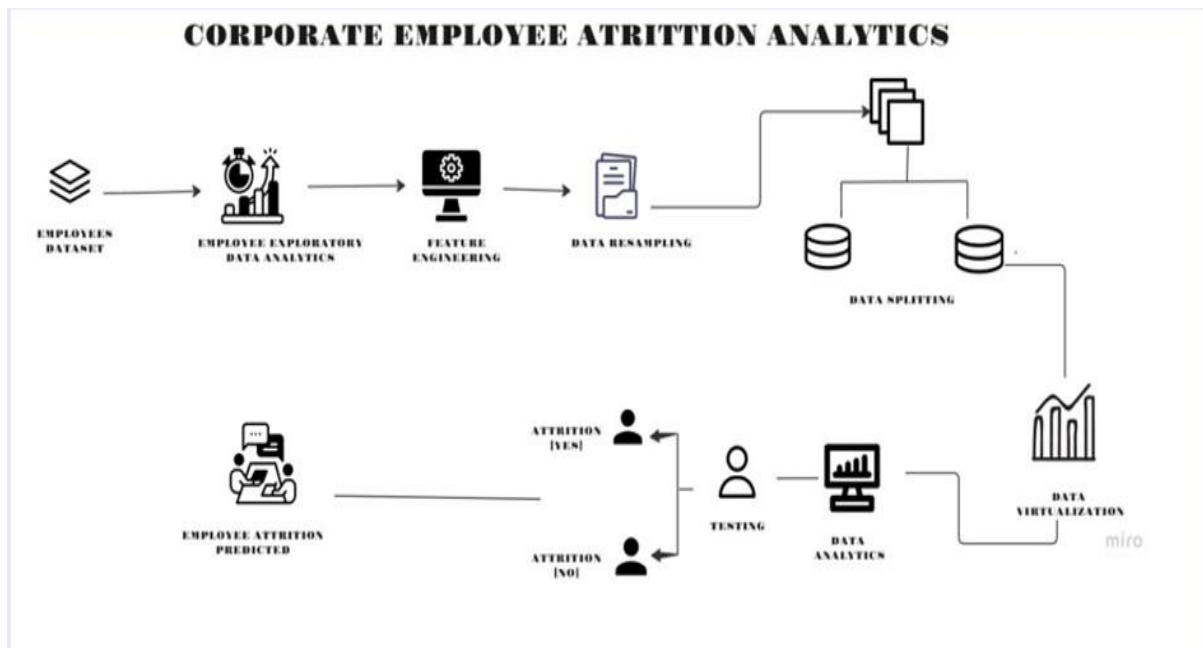
Let’s understand this by connecting this concept with two analogies. One is to wash freshly picked vegetables and second is only taking the wanted items to eat in the plate during the buffet. Washing of vegetables indicates the removal of dirt i.e. unwanted materials from the data. Here noise removal is done. Taking only eatable items in the plate is if we don’t need specific data then we should not consider it for further process. This whole process includes transformation, normalization etc.

#### **7.Modelling:**

Modelling decides whether the data prepared for processing is appropriate or requires more finishing and seasoning. This phase focuses on the building of predictive/descriptive models.

#### **8.Evaluation:**

Model evaluation is done during model development. It checks for the quality of the model to be assessed and also if it meets the business requirements. It undergoes diagnostic measure phase (the model works as intended and where are modifications required) and statistical significance testing phase (ensures about proper data handling and interpretation).



## 9. Deployment:

As the model is effectively evaluated it is made ready for deployment in the business market. Deployment phase checks how much the model can withstand in the external environment and perform superiorly as compared to others.

## 10. Feedback:

Feedback is the necessary purpose which helps in refining the model and accessing its performance and impact. Steps involved in feedback define the review process, track the record, measure effectiveness and review with refining.

# VISUALIZATIONS AND MODEL EVALUATION

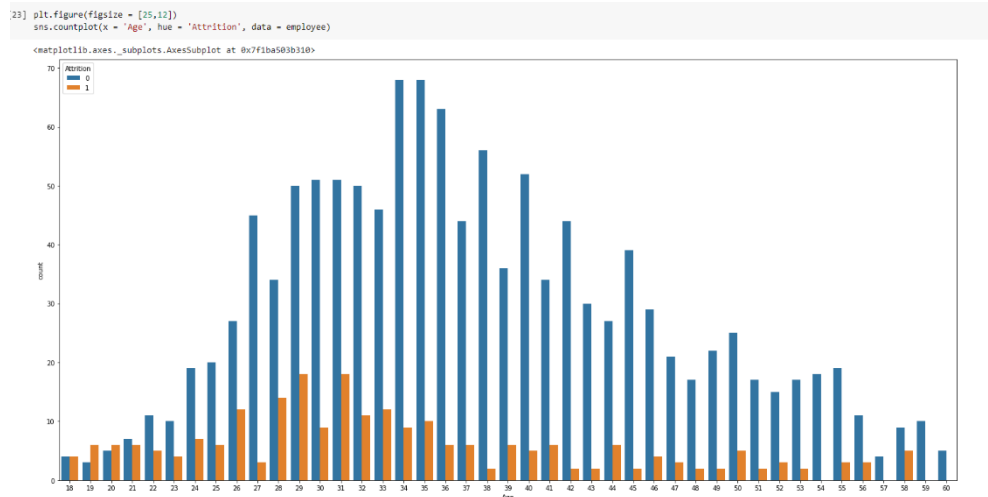
This figure shows the data frame (which isn't cleaned andsanitized as it has lots of null values)

employee

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	...	RelationshipSatisfaction	StandardHours
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	...	1	80
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	...	4	80
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	...	2	80
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	...	3	80
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	...	4	80
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1465	36	No	Travel_Frequently	884	Research & Development	23	2	Medical	1	2061	...	3	80
1466	39	No	Travel_Rarely	613	Research & Development	6	1	Medical	1	2062	...	1	80
1467	27	No	Travel_Rarely	155	Research & Development	4	3	Life Sciences	1	2064	...	2	80
1468	49	No	Travel_Frequently	1023	Sales	2	3	Medical	1	2065	...	4	80
1469	34	No	Travel_Rarely	628	Research & Development	8	3	Medical	1	2068	...	1	80

1470 rows × 35 columns

This figure shows the distribution of target variables based on age of employees

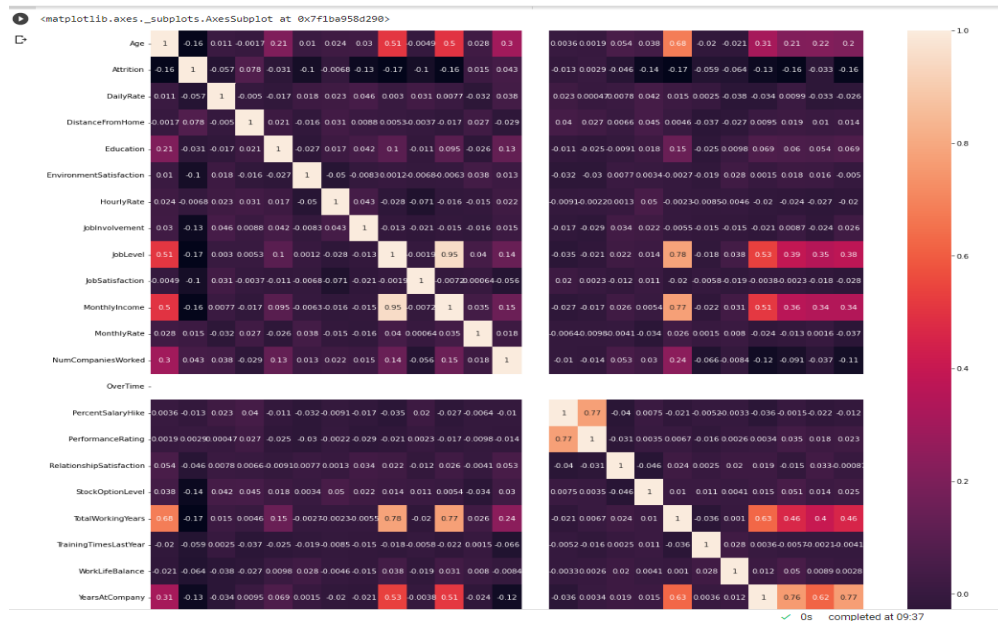




## The histogram plot of dataset

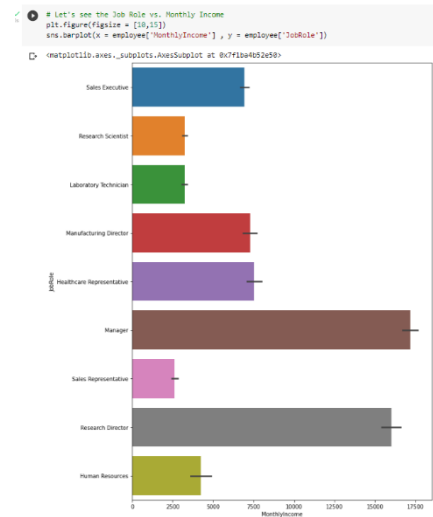
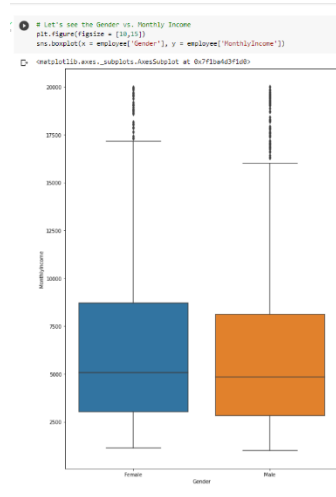
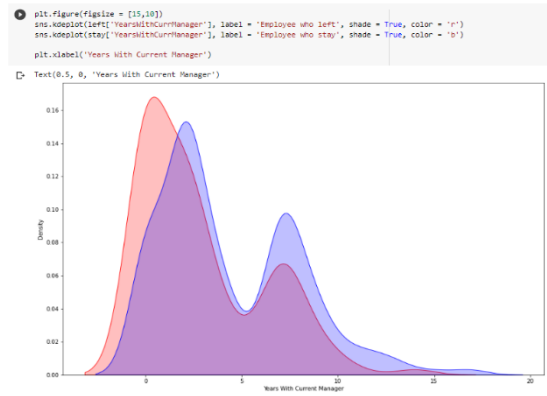
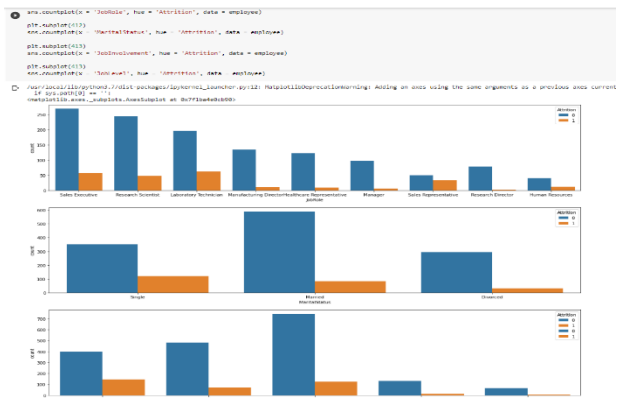


## Heat Map of Dataset



And then, we train the model and test for unknown data to predict the attrition rate. And the screenshot is attached herewith

# Visualizations done for checking relations between the attributes



# Model

## Train and Evaluate Logistic Regression Classifier

```
[40] from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25)

[41] X_train.shape
(1102, 50)

[42] X_test.shape
(368, 50)

[43] from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score

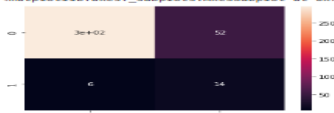
[44] model = LogisticRegression()
model.fit(X_train, y_train)

[45] y_pred = model.predict(X_test)

[47] from sklearn.metrics import confusion_matrix, classification_report
print('Accuracy {} %'.format(100 * accuracy_score(y_pred, y_test)))
Accuracy 84.23913043478261 %

[48] cm = confusion_matrix(y_pred, y_test)
sns.heatmap(cm, annot = True)

<matplotlib.axes._subplots.AxesSubplot at 0x7f1ba3d2f450>
```



```
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.85	0.98	0.91	302
1	0.70	0.21	0.33	66
accuracy			0.84	368
macro avg	0.78	0.60	0.62	368
weighted avg	0.82	0.84	0.81	368


## Train and Evaluate A Random Forest Classifier

```
[50] from sklearn.ensemble import RandomForestClassifier
model = RandomForestClassifier()
model.fit(X_train, y_train)

RandomForestClassifier()

# Testing Set Performance
cm = confusion_matrix(y_pred, y_test)
sns.heatmap(cm, annot = True)

<matplotlib.axes._subplots.AxesSubplot at 0x7f1ba35ff208>
```



```
[53] print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.84	0.99	0.91	302
1	0.77	0.15	0.25	66
accuracy			0.84	368
macro avg	0.81	0.57	0.58	368
weighted avg	0.83	0.84	0.79	368

## Train and Evaluate A Deep Learning Model

```
[54] import tensorflow as tf
model = tf.keras.models.Sequential()
model.add(tf.keras.layers.Dense(units = 500, activation = 'relu', input_shape = (50,)))
model.add(tf.keras.layers.Dense(units = 500, activation = 'relu'))
model.add(tf.keras.layers.Dense(units = 500, activation = 'relu'))
model.add(tf.keras.layers.Dense(units = 5, activation = 'sigmoid'))

[55] model.summary()
```

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 500)	25100
dense_1 (Dense)	(None, 500)	250500
dense_2 (Dense)	(None, 500)	250500
dense_3 (Dense)	(None, 5)	501

Total params: 527,001  
Trainable params: 525,000  
Non-trainable params: 0

```
model.compile(optimizer = 'Adam', loss = 'binary_crossentropy', metrics = ['accuracy'])
epoch_0_val = model.fit(X_train, y_train, epochs = 100, batch_size = 50)
```

Epoch	Loss	Accuracy
0/100	1.386e-05	1.0000
1/100	1.140e-05	1.0000
2/100	1.140e-05	1.0000
3/100	1.140e-05	1.0000
4/100	1.140e-05	1.0000
5/100	1.140e-05	1.0000
6/100	1.140e-05	1.0000
7/100	1.140e-05	1.0000
8/100	1.140e-05	1.0000
9/100	1.140e-05	1.0000
10/100	1.140e-05	1.0000
11/100	1.140e-05	1.0000
12/100	1.140e-05	1.0000
13/100	1.140e-05	1.0000
14/100	1.140e-05	1.0000
15/100	1.140e-05	1.0000
16/100	1.140e-05	1.0000
17/100	1.140e-05	1.0000
18/100	1.140e-05	1.0000
19/100	1.140e-05	1.0000
20/100	1.140e-05	1.0000
21/100	1.140e-05	1.0000
22/100	1.140e-05	1.0000
23/100	1.140e-05	1.0000
24/100	1.140e-05	1.0000
25/100	1.140e-05	1.0000
26/100	1.140e-05	1.0000
27/100	1.140e-05	1.0000
28/100	1.140e-05	1.0000
29/100	1.140e-05	1.0000
30/100	1.140e-05	1.0000
31/100	1.140e-05	1.0000
32/100	1.140e-05	1.0000
33/100	1.140e-05	1.0000
34/100	1.140e-05	1.0000
35/100	1.140e-05	1.0000
36/100	1.140e-05	1.0000
37/100	1.140e-05	1.0000
38/100	1.140e-05	1.0000
39/100	1.140e-05	1.0000
40/100	1.140e-05	1.0000
41/100	1.140e-05	1.0000
42/100	1.140e-05	1.0000
43/100	1.140e-05	1.0000
44/100	1.140e-05	1.0000
45/100	1.140e-05	1.0000
46/100	1.140e-05	1.0000
47/100	1.140e-05	1.0000
48/100	1.140e-05	1.0000
49/100	1.140e-05	1.0000



## **REFLECTIONS ON THE PROJECT**

Our Nalaiya Thiran project in collaboration with IBM and Tamil Nadu government has been the most rewarding and learning experiences we have had. With such empathetic, compassionate, and supportive mentors, this experience has helped me achieve my goal of completing my project. Because of the techniques we learned not only from my mentors and professors but from internet and books too.

We are confident that we will continue to grow and develop professionally and in my personal endeavors. Within my internship, there were two distinct learning experiences that stand out to me as the most influential aspects of my development this semester: community involvement in discussion forum and self-learning.

Throughout my project experience, we were able to develop and foster a truly positive and compassionate learning cum implementation environment, all through the support and mentorship of our mentors.

Through the application of time management, organization, discipline and consistent practice, our self-exploration and learning skills improved greatly. Additionally, my development both with the project we were given with and planning and implementing the same directly impacted our academic gain.

We are confident in our growth and development. We would not have the knowledge or skills we have today if it were not for our project experience with the industry mentor, college mentors and fellow interns.

## CONCLUSION

Overall, this project was a useful experience. We have gained new knowledge and skills we achieved several of my learning goals. We got insight into professional practice. We learned the different facets of working. We experienced that self-exploration, as in many organizations, is an important factor for the progress of projects. Related to our study we learned more about employee attrition rate prediction and the various approaches and algorithms to achieve the same. Furthermore, we have experienced that it is of importance of each strategy and how other one is better than the current algorithm and in which application. we found that the internship is not one sided, but it is a way of sharing knowledge, ideas and opinions and implementing the same to get results. The internship was also good to find out what our strengths and weaknesses are. This helped me to define what skills and knowledge. We believe that our time spent in learning and surfing regarding various algorithms and the mathematics behind was well worth it and contributed to finding an acceptable solution to build a model and predict the employee's attrition rate. Two main things that we've learned the importance of time-management skills and self- motivation. At last, this project has given us new insights and motivation to pursue a career in machine learning domain.

## **Link to code and executable file**

**<https://github.com/IBM-EPBL/IBM-Project-12732-1659460164>**