**PROJECT REPORT**

**Employee Attrition Analytics**

*Submitted towards the partial fulfillment of the criteria for award of KPMG Data Science Prodegree by Imarticus*

***Submitted By:***

*1. Jessica Sthul*

*2. Purva Goyal*

*3. Shelly Sharma*

*4. Akhilesh Sonarkar*

*5. Prasad Shinde*

***Course and Batch****: Data Science Prodegree-DSP 33*



**Acknowledgements**

We are using this opportunity to express my gratitude to everyone who supported us throughout the course of this group project. We are thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

Further, we were fortunate to have \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ as our mentor. She has readily shared his immense knowledge in data analytics and guide us in a manner that the outcome resulted in enhancing our data skills.

We wish to thank, all the faculties, as this project utilized knowledge gained from every course that formed the DSP program.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

**Certificate of Completion**

I hereby certify that the project titled “Employee Attrition Analytics” was undertaken and completed under my supervision by Group 2 from the batch of Data Science Prodegree DSP-33

Mentor:

Date: 20-03-2021

Place: Mumbai

**Table of contents**

1. **INTRODUCTION**
2. **Title &objective of the study**
3. **Need for the study**
4. **Data sources**
5. **Tools and techniques**
6. **DATA PREPARATION AND UNDERSTANDING**
7. **Data extraction and cleaning**
8. **Feature engineering**
9. **Data dictionary**
10. **Exploratory Data Analysis(EDA)**
11. **FITTING MODELS TO DATA**
12. **Decision Tree**
13. **Random Forest**
14. **Bagging**
15. **Boosting**
16. **Key findings**
17. **Conclusions**
18. **References**

**ABSTRACT**

‘Employee Attrition Analytics ‘gives an overview of analysis for a consulting firm and allows understanding and identifying factors that help in retaining employees.

**INTRODUCTION**

**TITLE** - Employee Attrition Analytics

**OBJECTIVE-**

The objective of this project is to build a data model which will

* Identify the factors influencing attrition
* Predict possible attrition
* Identify possible ways to retain high performers.

**NEED FOR THE STUDY-**

* Employee attrition refers to the loss of employees through several circumstances, such as resignation and retirement.
* Each industry has its own standards for acceptable attrition rates. Due to the expenses associated with training new employees, any type of employee attrition is typically seen to have a monetary cost.
* Attrition rate at any organization for a given month is calculated as the total number of employees leaving the firm divided by the total headcount for that month.
* Studying such analysis helps the company to identify employees who have high possibility of quitting and will allow them to allocate resources to combat the issue.

**DATA SOURCES –**

* We were provided with the three datasets along with the problem statement.
* The two datasets contained records of the employees about different work hours and other details of the employees for two years 2016-2018.
* The third dataset ‘Terminations’ contained all the details about resigned employees and the factors related to attrition.

**TOOLS AND TECHNIQUES –**

* We are doing coding in Python using Jupiter notebook

**DATA PREPARATION AND UNDERSTANDING**

To get deeper understanding of data or to apply different models to the datasets we need to prepare our data first, so that we can extract the data to read, finding actual predictors and perform visualization part also. Here we performed few steps to get the full knowledge

We imported the essential libraries as below

* import pandas as pd
* import seaborn as sns
* import matplotlib.pyplot as plt
* import numpy as np
* from sklearn import tree
* from sklearn import preprocessing
* from sklearn.ensemble import AdaBoostClassifier # we are applying the boosting techinque
* from sklearn.tree import DecisionTreeClassifier
* from sklearn import metrics
* import matplotlib.pyplot as plt
* %matplotlib inline
* from sklearn.model\_selection import ShuffleSplit, train\_test\_split,cross\_val\_score,GridSearchCV,KFold, RandomizedSearchCV

**Setting the working directory :**

* We are setting the working directory to import the dataset.

**Read the datasets :**

* Now we are reading the datasets into object df1, df2 and df3.

**Concatenating the all datasets** :

* We are concatenating the df1, df2 and df3 using append command setting into object df.

**Dropping the variables :**

We are going to drop the unnecessary variables which don’t have any impact of the data. And we are going to store the data into new\_data object. So now our variables becomes 208.

**Summarizing data :**

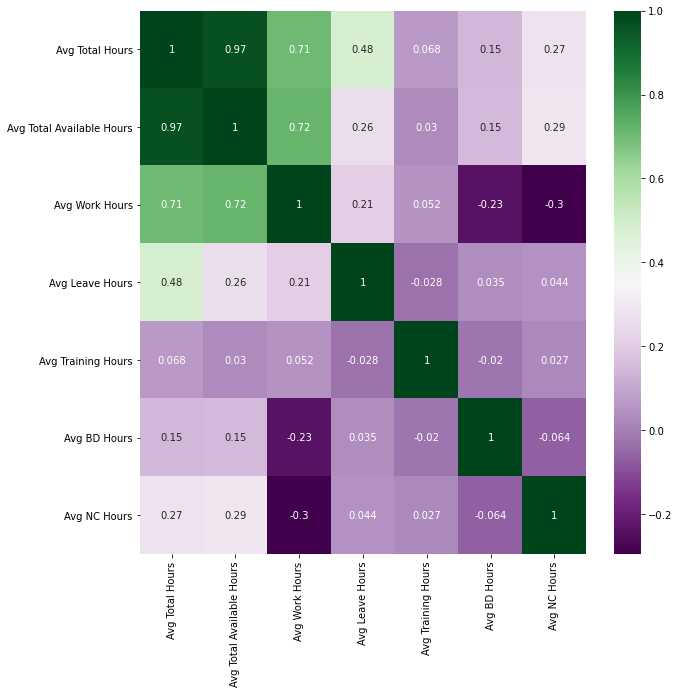
The **describe()** function computes a summary of statistics pertaining to the DataFrame columns.

This function gives the **mean, std** and **IQR** values. And, function excludes the character columns and given summary about numeric columns

**Correlation :**

There may be complex and unknown relationships between the variables in your dataset.

It is important to discover and quantify the degree to which variables in your dataset are dependent upon each other. This knowledge can help you better prepare your data to meet the expectations of machine learning algorithms, such as linear regression, whose performance will degrade with the presence of these interdependencies.



In our dataset between Avg Total Hours and Avg Total Available Hours, Avg Work Hours, Avg Training Hours showing positive correlation which occurs when large values of Avg Total Hours corresponding to large values of the others, and vice versa. And between Avg Leave Hours and Avg Training Hours, Avg Training Hours and Avg BD Hours, Avg Work Hours and Avg BD Hours has negative correlative correlation which occurs when large value of one features corresponding to  small values of the others, and vice versa.

**FEATURE ENGINEERING**

Feature engineering has two goals primarily:

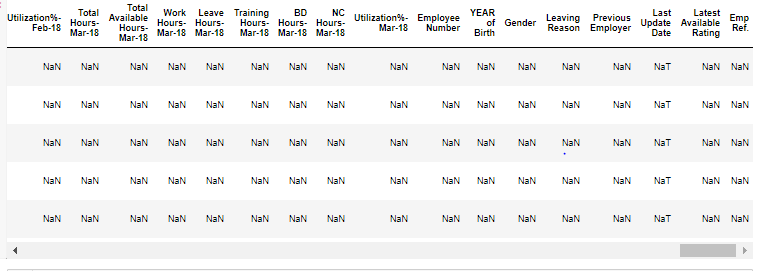
* Preparing the proper input dataset, compatible with the machine learning algorithm requirements
* Improving the performance of machine learning models

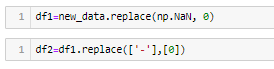
**List of Feature Engineering Techniques**

1. We are removing the columns which are not important for our employee attrition analysis. The columns that are removed are listed below

* Employee No
* Employee Number
* Join Date
* Supervisor name
* Previous Employer
* Last Update Date
* Emp Ref.
* Employee Name
* Latest Available Rating
* YEAR of Birth
* Termination Date

1. We have some garbage data which is available as ‘NaN’ and ‘-‘. We have replaced them as ‘zero’





**EXPLORATORY DATA ANALYSIS**

**DATA PREPARATION AND UNDERSTANDING**

1. **Data dictionary**

****

1. **Exploratory Data Analysis(EDA)**

## Univariant analysis – Visual Analysis - Distribution and barplots and boxplots etc.

### We plotted the different bar plots to analyze what type of data the dataset contains, for the below categorical columns

Profit Center,

Employee Position,

Employee Location,

People Group,

Employee Category,

Current Status,

Gender,

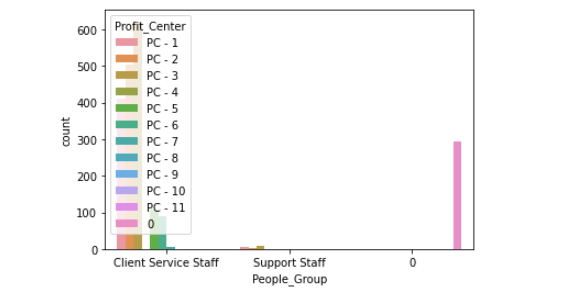
Leaving Reason

We plotted the different count plots to analyze the data which has the major impact on employee attrition

1. Plotted the 1st graph between 'People Group' and 'Profit Center'

and as per the below graph analyzed the below pointers

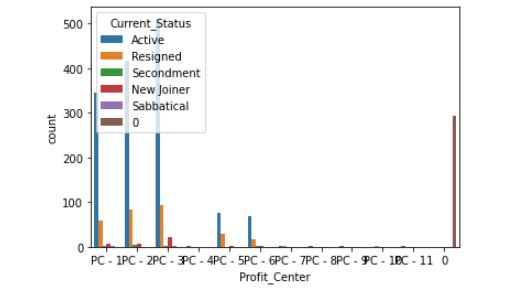
* Approximately 650 employees have client group and they are falls in profit center 3
* 300 employees don’t have any group



1. Plotted the 2nd graph between ‘Profit Center’ and 'Current Status'

and as per the below graph analyzed the below pointers

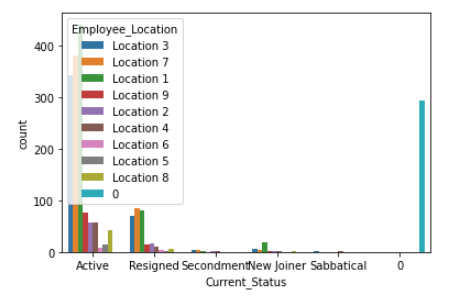
* In case of profit center 1 and 2 active employees are very high.
* profit center 6,7,8,9 there are no employees active



1. Plotted the 3rd graph between Current Status and Employee Location

and as per the below graph analyzed the below pointers

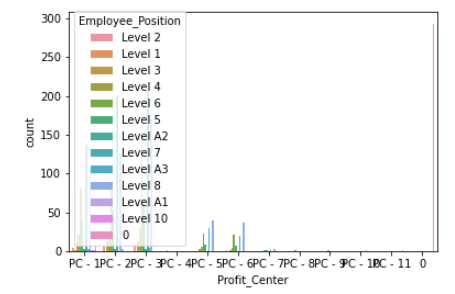
* Location 3,7 and 1 almost 400 employees are active and 100 employees resigned



1. Ploted the 4th graph between Current\_Status and Employee\_Position

and as per the below graph analysed the below pointers

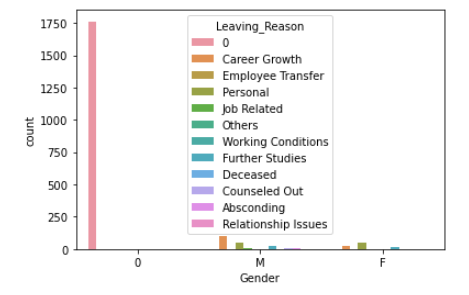
* large no. of employees which do not have any position
* profit center 1 there are 200 employees which falls in level 8 and 90 employees which falls in level 5



1. Ploted the 5th graph between Gender and Leaving\_Reason

and as per the below graph analysed the below pointers

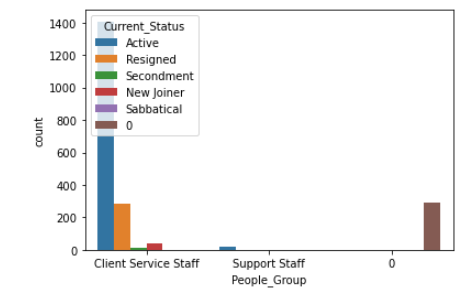
* There are 1750 employees which dont give leaving reason
* 100 males leaving company because of career growth
* Approximetly 50 females giving job related reasons



1. Ploted the 6th graph between People\_Group and Current\_Status

and as per the below graph analysed the below pointers

* In Client group 1400 employees are active and 240 employees resigned
* In Service Staff 30 employees are New Joiner
* There are approximetly 250 employees which do not follow any group



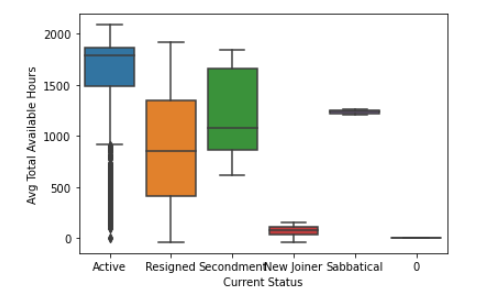
1. Plotted the 7th graph between People Group and Current Status

and as per the below graph analyzed the below pointers

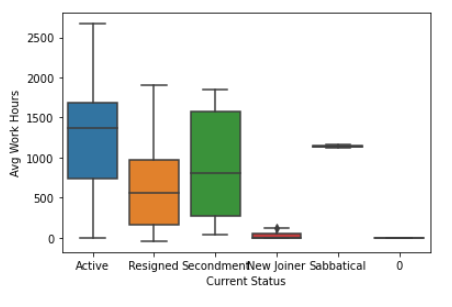
* In Client group 1400 employees are active and 240 employees resigned
* In Service Staff 30 employees are New Joiner
* There are approximately 250 employees which do not follow any group

We plotted various boxplots to get the bigger picture of the data since we have the correlation between different columns hence visualized the data with the help of boxplots

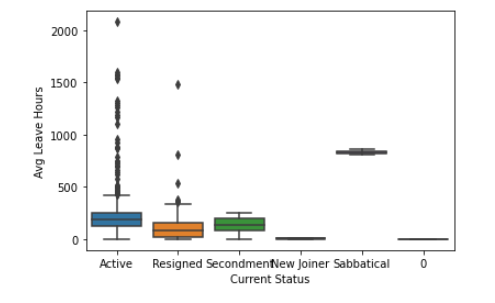
1. Plotted the 1st boxplot between Current Status and Avg Total Available Hours

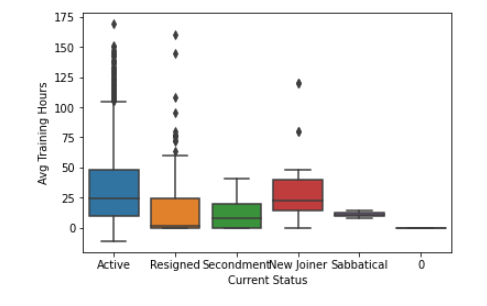


1. Plotted the 2nd plot between Current Status and Avg Work Hours



1. We plotted the 3rd plot between Current Status and Avg Leave Hours



1. We plotted the plot between Current Status and Avg Training Hours 

**FITTING MODELS TO DATA**

**DECISION TREE**

* We create a Decision tree which is our first classifier over here and we used the gini index.
* Decision tree provide an effective method of Decision Making because it clearly lay out the problem so that all options can be challenged.
* Provide a framework to quantify the values of outcomes and the probabilities of achieving them.
* Decision Tree gives us more than 98% accuracy.

**RANDOM FOREST**

* We applied Random Forest on the Training data set to validate if any further improvement of the model can be performed post the Decision Tree.
* Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.
* Random forest gives us more than 97% accuracy.

**BAGGING**

* Later on we applied bagging, in bagging we combine the outputs of multiple classifiers trained on different samples of the training data. This helps in reducing overall variance.
* Due to the reduction in variance, normally unstable classifiers can be made robust with the help of bagging.
* Bagging gives us more than 97% accuracy.

**BOOSTING**

* In the last we applied Boosting which is a method of merging different types of predictions.
* Boosting grants power to machine learning models to improve their accuracy of prediction
* Boosting decreases the bias error and builds strong predictive models
* Boosting gives us more than 98% accuracy.

**KEY FINDINGS**

The following table shows which of data models performed the best and the accuracy is given.

|  |  |  |
| --- | --- | --- |
| Algorithms |  | Accuracy |
| 1. Decision tree | 98.29% | |
| 1. Random forest | 97.81% | |
| 1. Bagging |  | 97.32% |
| 1. Xgboost | 98.54% | |

**CONCLUSION-**

* Using the provided datasets and data dictionary, we successfully extracted, visualized and analyzed the data, studied the parameters which affected the attrition and also predicted possible attritions with high accuracy.
* From the count plots showing employee positions and leaving reason, we can conclude that the maximum resigned employees falls in low level positions as well as their common leaving reasons is career growth.
* This factor could get rectified by providing good opportunities along with training for the higher posts for the aspiring employees. The given data also shows lesser training hours for resigned employees
* Another boxplot also shows the lesser leave hours for resigned people and high leave hours for active people, this might indicate employees might be resigning due to overworking, it indicates they could not find the work-life balance.
* To overcome this problem, company can assigned the leave hours to each employee, so that they are able to enhance their productivity eventually becoming good performers.

**REFERENCES-**

Study material provided by imarticus learning

* Imarticus online video lectures
* Imarticus online study material.