DATATRAINED ACADEMY-Blog Article -1

Submitted by: Sucharitha A M

Btatch:1829

## HR Analytics Project-Understanding the Attrition in HR

Problem Statement:

Write stories/blogs on any 2 projects in 2000 words.

Every year a lot of companies hire a number of employees. The companies invest time and money in training those employees, not just this but there are training programs within the companies for their existing employees as well. The aim of these programs is to increase the effectiveness of their employees. But where HR Analytics fit in this? and is it just about improving the performance of employees?

**HR Analytics**

Human resource analytics (HR analytics) is an area in the field of analytics that refers to applying analytic processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on investment. HR analytics does not just deal with gathering data on employee efficiency. Instead, **it aims to provide insight into each process by gathering data and then using it to make relevant decisions about how to improve these processes.**

**Attrition in HR**

Attrition in human resources refers to the gradual loss of employee’s overtime. In general, relatively high attrition is problematic for companies. HR professionals often assume a leadership role in designing company compensation programs, work culture, and motivation systems that help the organization retain top employees.

How does Attrition affect companies? and how does HR Analytics help in analysing attrition? We will discuss the first question here and for the second question, we will write the code and try to understand the process step by step.

**Attrition affecting Companies**

A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork, and new hire training are some of the common expenses of losing employees and replacing them. Additionally, regular employee turnover prohibits your organization from increasing its collective knowledge base and experience over time. This is especially concerning if your business is customer-facing, as customers often prefer to interact with familiar people. Errors and issues are more likely if you constantly have new workers.

## 1.Problem Definition:

Analysing the HR Criteria of a Company and how they promote their Employees and keep Balance between them using Data Analytics, Data Visualizations, and Machine Learning Models for Classification Purposes.HR leaders must align HR data and initiatives to the organization’s strategic goals. For example, a tech company may want to improve collaboration across departments to increase the number of innovative ideas built into their software. HR initiatives like shared workspaces, company events, collaborative tools, and employee challenges can be implemented to achieve this goal. To determine how successful initiatives are, HR analytics can be utilized to examine correlations between initiatives and strategic goals. Once data is gathered, HR analysts feed workforce data into sophisticated data models, algorithms, and tools to gain actionable insights. These tools provide insights in the form of dashboards, visualizations, and reports. An ongoing process should be put in place to ensure continued improvement. Benchmark analysis Data-gathering Data-cleansing Analysis Evaluate goals and KPIs Create action plan based on analysis (continuously test new ideas) Execute on plan Streamline process.

## 2. Data Analysis:

The key to success in an organisation is the ability to attract and retain top talents. It is vital for the Human Resource (HR) Department to identify the factors that keep employees and those who the first stage of this analysis is to describe the dataset, understand the meaning of each variable, detect possible patterns and perform the necessary adjustments to ensure that the data will be proceeded correctly during the Machine Learning process. each prompt them to leave. Organisations could do more to prevent the loss of good people.



This project is based on a hypothetical dataset downloaded from  [HR Analytics Employee Attrition & Performance](https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset). It has 1,470 data points (rows) and 35 features (columns) describing each employee’s background and characteristics; and labelled (supervised learning) with whether they are still in the company or whether they have gone to work somewhere else. Machine Learning models can help to understand and determine how these factors relate to workforce attrition

# Data Preparation and Cleaning

* Reading the CSV file and doing initial statistical analysis (shape, values etc)
* Data Pre-processing: Reading the uniques values for each column and removing those which won’t be significant in the analysis further.
* Create a new data frame to proceed with the analysis further

Dataset contains

Age

Attrition

Business Travel

Daily Rate

Department

DistanceFromHome

Education

Education Field

Employee Count

Employee Number

Environment Satisfaction

Gender

Hourly Rate

Job Involvement

Job Level

Job Role

Job Satisfaction

Marital Status

Monthly Income

Monthly Rate

NumCompaniesWorked

Over18

Overtime

PercentSalaryHike

Performance Rating

Relationship Satisfaction

Standard Hours

StockOptionLevel

TotalWorkingYears

TrainingTimesLastYear

WorkLifeBalance

YearsAtCompany

YearsInCurrentRole

YearsSinceLastPromotion

YearsWithCurrManager

Analyse Dataset:

Let’s import the relevant Python libraries, and read in the data file. Jupyter notebook with Python codes

df=pd.read\_csv("HR\_Employee\_Attrition.csv")

df.head()

The correlation between different features of the dataset showed that employees with low satisfaction level left. The correlation heat map is shown below:

plt.figure(figsize=(10,6))

sns.heatmap(dfcor,cmap='Oranges',annot=**True**)

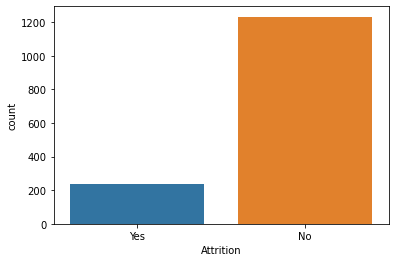


The correlation matrix does not indicate any high degree of correlation with the dependent variable. However, it does provide us with a holistic view off all the factors.

## 3. EDA Concluding Remark

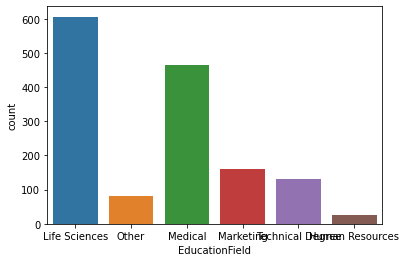
* Find patterns of data through visualization and reveal the hidden trends from data.
* Using both matplotlib and seaborn library to visualize the data
* Finding relationships between features using bar graphs, histograms, box plots, heatmap
* Analyzing both the numerical and the categorical columns separately

Here Attrition will be the target variable. The dataset is well organised with no missing values Target class is imbalance, with attrition rate of 16%.



Are employees leaving because they are poorly paid. Employees are paid an hourly rate of $30 to $100, and attrition seems to happen at every level regardless of employee hourly rate. This can be confirmed later at feature importance.

Education Field seems to be one of the key factors to attrition, as a larger proportion of education field employees has departed.

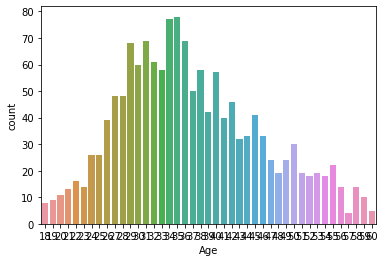


There are only 3 departments included for this analysis.

The visualization of the EDA process:

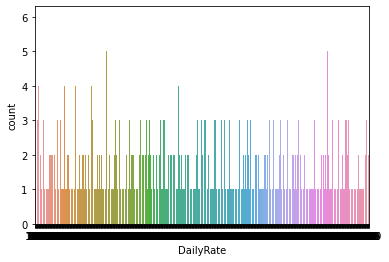
sns.countplot(df['Age']);

Age seems to be one of the key factors to attrition, as a larger proportion of Age employees has departed.

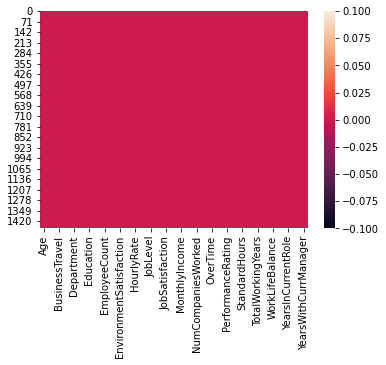


sns.countplot(df['Daily Rate'])

compare to the Attrition and education field plot at Daily Rate will be little more different plot.



sns.heatmap(df.isnull())



Heatmap contains the plot null values of the dataset.

## 4. Pre-Processing Pipeline:

For the model to proceed with the data efficiently, the categorical variables salary and department have been encoded. As the values of salary have an order, they have been encoded into integers within the same variable. For department, as the values have no specific order, they have been encoded into individual variables with Boolean values. Thus, the dataset has been transformed from 10 variables to 19 variables. Numerical variables scaled between 0 and 1 to remove any influence of their difference in value ranges on the model. They have also been checked for skewness, without a real change on their shape.

Data has to be pre-processed as machine learning models are better at reading numbers than words. Using label encoding, categorical data can be replaced with numbers. Below code is to display all categorical data.

**from** **sklearn.preprocessing** **import** Label Encoder

LE=Label Encoder()

df["Attrition"]=LE.fit\_transform(df["Attrition"])

using the above label encoding method, categorical data can be replaced with number.

## 5. Building Machine Learning Models:

As the dataset is imbalance, use cross validation when training the models, and each baseline model performance can be tabulated.

The model will be cross-validated using a 10-fold cross validation method returning the average accuracy. This method will be applied at every modelling step, to ensure that the model is not biased by the training set split.

x\_train = df.drop(['Attrition'],axis=1)

y\_train = df.Attrition

x\_train, x\_test, y\_train, y\_test=train\_test\_split(x, y, random\_state=0, test\_size=0.2)

Here we using the classification Methods to build the models:

There are 6 classification models

1. DecisionTreeClassifier

2 Logistics Regression

3 KNeighborsClassifier

4 SVM Classifier

5 Navie byes Classifier

6 Random Forest Classifier

The baseline model performance results are terrible, with F1-scores ranging from 20% to 40% for most models. After tuning hyperparameters and the threshold, the Logistic Regression has achieved F1-score 56.0% and **Recall 57.0%**.

The baseline model performance results are terrible, with F1-scores ranging from 20% to 40% for most models. After tuning hyperparameters and the threshold, the DecisionTreeClassifier has achieved F1-score 59.0% and **Recall 60.0%**.

The baseline model performance results are terrible, with F1-scores ranging from 20% to 40% for most models. After tuning hyperparameters and the threshold, the KNeighborsClassifier has achieved F1-score 49.0% and **Recall 51.0%**.

The baseline model performance results are terrible, with F1-scores ranging from 20% to 40% for most models. After tuning hyperparameters and the threshold, the SVM Classifier

has achieved F1-score 46.0% and **Recall 50.0%**.

The baseline model performance results are terrible, with F1-scores ranging from 20% to 40% for most models. After tuning hyperparameters and the threshold, the Navie byes Classifier has achieved F1-score 70.0% and **Recall 74.0%**.

The baseline model performance results are terrible, with F1-scores ranging from 20% to 40% for most models. After tuning hyperparameters and the threshold, the Random Forest classifi

r has achieved F1-score 62.0% and **Recall 60.0%**.

Also, if features are closely related to one another .one of them has to be removed to prevent misleading results to linear models such as Logistic Regression. Although tree-based models are not directed affected, they could also lead to over-fitting.

The highest final model of the dataset: In the end, we can see that utilizing data science on employee's attrition provided significant benefit to the business as we can tag each employee with the attrition score and come up with customized HR retention strategy for each group

lr = Logistic Regression()

lr.fit(x\_train,y\_train)

y\_pred\_lr = lr. Predict(x\_test)

lr\_roc\_auc = roc\_auc\_score(y\_test, lr. Predict(x\_test))

fpr,tpr,thersholds = roc\_curve(y\_test,lr.predict\_proba(x\_test)[:,1])

rf = RandomForestClassifier()

rf.fit(x\_train,y\_train)

y\_pred\_rf = rf.predict(x\_test)

rf\_roc\_auc = roc\_auc\_score(y\_test, rf.predict(x\_test))

fpr,tpr,thersholds = roc\_curve(y\_test,rf.predict\_proba(x\_test)[:,1])

plt.figure()

plt.plot(fpr,tpr,label='Logistic Regression' %lr\_roc\_auc)

plt.plot(fpr,tpr,label='RandomForestClassifier'%rf\_roc\_auc)

plt.plot([0,1],[0,1],'r--')

plt.xlim([0.0,1.0])

plt.ylim([0.0,1.05])

plt.xlabel('False Positive Rate')

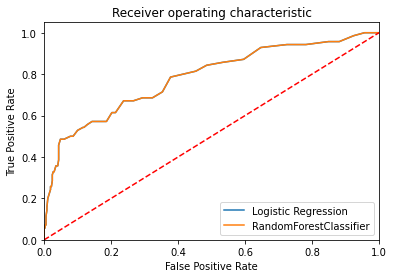
plt.ylabel('True Positive Rate')

plt.title('Receiver operating characteristic')

plt.legend(loc="lower right")

plt.savefig('ROC')

plt.show()



According to the classification report the accuracy of the model is 87% however its recall is lower at 43% of positive cases. The RandomForestClassifiermodel is providing excellent results, however the purpose of the problem is to identify employees that are likely to leave. This is the reason that recall then becomes a very important measure. Recall measures the fraction of values that are identified correctly.

Random Forest Classifier has emerged as the final winning model with F1-score 62.0% and highest **Recall 60.0%**. This could be the highest possible score achieved with the inherent limitations in the dataset.

The top factor for employee attrition in this hypothetical organisation seems to be **monetary**, emerged at the top. This could be due to a bad compensation process or causing a poor work-life balance. The next important factor seems to be **personal relationships**with follow workers, where current manager and job role could be the main contributing reasons for attrition. Finally, **employee engagement**is a critical satisfaction factor, and the organisation should keep employees constantly involved and motivated.

Machine learning models are as good as the data to feed it, and more data would strengthen the model. For example, in this dataset, the feature ‘Performance Rating’ has been restricted to scores of 3 and 4 only. More insights could be generated if the full spectrum of performance ratings is included. In the real-life situation, getting the right data is often more challenging than the analytics itself.

## 6. Concluding Remarks.

HR Analytics is gaining traction in organisations that embrace digital transformation. The scope has expanded from analytics of employee work performance to providing insights so that decisive improvements can be made to organisational processes. While some level of attrition is inevitable, it should be kept at the minimal possible level.

This model will allow the company to calculate the probability of an employee to leave the company and to act on key-factors to avoid departures. The satisfaction of employees and the amount of workload they have to bear seem to be important causes of withdrawals. A particular attention on the work-life balance would be crucial to improve the turnover rate.