**HR Analytics**

**A person looking at a person's profile

Description automatically generated**

1. **Problem Definition**

In human resources, attrition is the rate at which employees leave a company and is a key metric for understanding employee retention.  The true cost of replacing an employee can often be quite large. Job postings, hiring processes, paperwork, and new hire training are some of the common expenses of losing employees and replacing them. Top talent, in particular, can be very difficult and expensive to replace. The more talented the worker, the greater the consequences of attrition. Replacing an individual employee typically costs [one-half to two times](https://www.gallup.com/workplace/247391/fixable-problem-costs-businesses-trillion.aspx#:~:text=The%20cost%20of%20replacing%20an,to%20%242.6%20million%20per%20year.) the worker’s annual salary. Additionally, regular employee turnover prohibits your organization from increasing its collective knowledge base and experience over time.

Understanding why and when employees are most likely to leave can lead to actions to improve employee retention as well as possibly planning new hiring in advance. We will be using a step-by-step systematic approach that could be used for a variety of Machine Learning problems. This project would fall under what is commonly known as **HR Analytics** or **People Analytics**.

The project involves building up a predictive model to determine whether employees are likely to leave or stay within the company. The main objective is to assist the HR department to reduce the attrition rates and implement effective strategic plans to increase the retention rate.

* What is the **likelihood** of an active employee leaving the company?
* What are the key **indicators** of an employee leaving the company?
* What **strategies** can be adopted based on the results to improve employee retention?

In addressing the ongoing challenges of the pandemic and the rise of remote work, employee attrition analytics will remain important to organizations seeking to retain top talent. Predictive analytics (HR Analytics) capability enables the design of an employee retention model to keep these valuable employees engaged and on board.

1. **Data Analysis**

In this case study, a HR dataset was sourced from [IBM HR Analytics Employee Attrition & Performance](https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/" \t "_blank) which contains employee data for 1,470 employees with various information about the employees. We will use this dataset to predict when employees are going to quit by understanding the main drivers of attrition.

Objective of the Data Analysis: The primary goal is to analyse the features and targets influencing employee attrition and its financial and operational impacts on the company.

* 1. **Data Description**
* First, we import the dataset and make a copy of the source file for this analysis.
* The dataset contains 1,470 rows and 35 columns.
* The dataset contains several numerical and categorical columns providing various information on employees personal as well as employment details.
* The columns are mainly of three data types (int64, float64 and object)
* The variables and their description:

Age: Employee Age

Attrition : Employee leaving the company (0=No, 1=Yes)

BusinessTravel : (1=Non-Travel, 2=Travel\_Frequently, 3=Travel\_Rarely)

DailyRate : Numerical Value - Salary Level

Department : (1=Human Resources, 2=Research & Development, 3=Sales)

DistanceFromHome : Numerical Value

Education : Numerical Value

'Below College' ,'College',’Bachelor’,’Master’,’Doctrate’,

EducationField : (1=Human Resources, 2=Life Sciences, 3=Marketing, 4=Medical, 5=Other, 6=Technical Degree)

EmployeeCount : Numerical Value

EmployeeNumber : EMPLOYEE ID

EnvironmentSatisfaction : Numerical Value : 1 'Low' 2 'Medium' 3 'High' 4 'Very High'

Gender : (1=Female, 2=Male)

HourlyRate : Numerical Value

JobInvolvement : Numerical Value : 1 'Low' 2 'Medium' 3 'High' 4 'Very High'

JobLevel : Numerical Value

JobRole : (1=Healthcare Representative, 2=Human Resources, 3=Laboratory Technician, 4=Manager, 5=Manufacturing Director, 6=Research Director, 7= Research Scientist, 8=Sales Executive, 9=Sales Representative)

JobSatisfaction : Numerical Value

MaritalStatus : (1=Divorced, 2=Married, 3=Single)

MonthlyIncome : Numerical Value

MonthlyRate : Numerical Value

NumCompaniesWorked : Numerical Value

Over18 : (1=Yes, 2=No)

OverTime : (1=No, 2=Yes)

PercentSalaryHike : Numerical Value – Percentage Increase In Salary

Performance Rating: Numerical Value : 1 'Low' 2 'Good' 3 'Excellent' 4 'Outstanding'

RelationshipSatisfaction : Numerical Value

StandardHours : Numerical Value

StockOptionLevel : Numerical Value

TotalWorkingYears : Numerical Value

TrainingTimesLastYear : Numerical Value - HOURS SPENT TRAINING

WorkLifeBalance : Numerical Value : 1 'Bad' 2 'Good' 3 'Better' 4 'Best'

YearsAtCompany : Numerical

YearsInCurrentRole : Numerical Value

YearsSinceLastPromotion : Numerical Value

YearsWithCurrManage : Numerical Value – Years working with current manager.

* 1. **Data Cleaning**
* Missing Data Handling: There were no missing values in the dataset.
* Dropping Values: The Employee Count, Over18 and Standard Hours have only 1 count so they can be dropped as they will not make much impact in our model. Also, Employee Number is taken on the basis of unique ID of the employees which does not helps so we can drop this too.
* Duplicate Removal: There were no duplicate records in the dataset.
  1. **Visualisation Analysis based on the statistical summary**
* The counts of all the columns are same which means there are no missing values present in the data.
* The mean is more than the median (50%) in most of the columns which means they are skewed to right.
* The min age of the employee is 18 and max is 60 and most of the employees are in between 36.
* In few columns the median (50%) is more than the mean which means they are skewed to left.
* Daily Rate has a huge outlier from the maximum of 1499 which is quite far from the 75% percentile. These differences we can also be seen in Age, DistanceFromHome, HourlyRate, Monthly Income, TotalWorkingYears, YearsAtCompany and many other.
* Some of the columns have huge difference in mean and the standard deviation.

A collage of graphs

Description automatically generated

* 1. **Visualisation based on feature and target**
* Univariate Analysis was done on independent variables to see their count.
* Bivariate Analysis was done between target and features so that we could understand the relationship.
* Multivariate Analysis was also done.
* We were able to see the outliers present in the dataset with box plot.
* We were able to see the positive and negative relationship between features and the target.
  1. **Outlier Handling**
* Outliers were present in the columns MonthlyIncome, NumCompaniesWorked, performanceRating, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion and YearsWithCurrManager. We will remove these outliers using Zscore method.
* Our dataset dimension after the removal of outliers was 1387 rows and 31 columns.
  1. **Encoding Variables**
* Machine Learning algorithms can typically only have numerical values as their predictor variables. Hence **Label Encoding** becomes necessary as they encode categorical labels with numerical values.
* We used Label Encoder for encoding the variables - 'Attrition', 'BusinessTravel', 'Department', 'EducationField', 'Gender', 'Job Role', 'MaritalStatus', 'OverTime'.
  1. **Skewness Analysis**
* We were able to see the skewness during our visual representation and were able to interpret the statistical information which we got from the dataset statistics summary. Next based on the skewness we got to check for the skewness correction.
* We used Yeo Johnson method for removing the skewness, after which we were able to get normal distribution for the variables.
  1. **Multicollinearity Analysis and Feature Selection**
* Correlation was perfectly visualised through heat map.
* This heatmap contains both positive and negative correlations.
* The target has more correlation with DistanceFromHome(+0.08)
* The target is negative correlated with TotalWorkingYears (-0.17)
* The column Job Level is correlated with MonthlyIncome (+0.9), and TotalWorkingYears (+0.7)
* The column Age is positive correlated with TotalWorkingYears (+0.6)
* The columns YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager, and TotalWorkingYears are correlated with each other.
* We don’t see much multicollinearity in the data.

A blue and green squares with black text

Description automatically generated with medium confidence

* We observed that business travel and hourly rate doesn’t have relationship with the target. We will be dropping the two columns.
  1. **Balancing the dataset (Target)**
* Machine learning algorithms typically work best when the number of instances of each class are roughly equal. We will have to address this target feature imbalance prior to implementing our Machine Learning algorithms.
* We will be using the SMOTE for balancing the target.

1. **EDA Concluding Remarks**

* The dataset does not feature any **missing or erroneous data values**, and all features are of the correct data type.
* The strongest **positive correlations** with the target features are Performance Rating, Monthly Rate, Num Companies Worked, Distance From Home.
* The strongest **negative correlations** with the target features are: Total Working Years, Job Level, Years in Current Role, and Monthly Income.
* The dataset is **imbalanced**
* **Higher age groups** showed lower Attrition Rate.
* We can observe that attrition rate is more in **Research and Development department**
* We can observe highest attrition is seen in the level of employees who have completed **Bachelor Education.**
* We can observe that attrition level is significant in employees who are from **life science education background.**
* We can observe that employees with **low daily rate** want to leave the organization when compared to high daily rate.
* **Single employees** show the largest proportion of leavers, compared to Married and Divorced counterparts.
* About 10% of leavers left when they reach their **2-year anniversary** at the company.
* People who **live closer to their work** show higher proportion of leavers compared to their counterparts.
* We can observe there is more of attrition for **the monthly salary below 5000**.
* We can observe that the employees have **low level of job involvement and job satisfaction** are leaving the organisation.
* We can observe that female employees with **work life balance** issue face higher attrition rate compared to male employees.
* People who must work **overtime**show higher proportion of leavers compared to their counterparts.
* Employees that have already worked at several companies previously show higher percentage of leavers compared to their counterparts.

1. **Pre-processing Pipeline**

In this part, we undertake data pre-processing steps to prepare the datasets for Machine Learning algorithm implementation.

**4.1 Feature Scaling**

* Feature Scaling is done using standard scalar which is a preprocessing technique for standardizing features by removing the mean and scaling to unit variance.

scaler = StandardScaler ()

x = pd. DataFrame(scaler.fit\_transform(x), columns=x. columns

**4.2 Splitting Data into training and testing set**

* The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.

#Creating Train Test Split

* x\_train, x\_test, y\_train, y\_test = train\_test\_split (x, y, test\_size=0.3, random\_state=random\_state)
* We have split the dataset into two sets, and we would be using them in building our machine learning models.

**5.Building Machine Learning Models**

* We would be working on classification Algorithms for bulding up our model and testing it. Based on the accuracy score and cross validation score, we would be choosing which model has better performance and accuracy rate for our dataset.
* The models which we went ahead were
  + Logistics Regression classifier
  + Decision Tree Classifier
  + Random Forest Classifier
  + Support Vector Machine Classifier
  + KNN Classifier
  + ADA Boost Classifier

Based on accuracy score, confusion matrix, classification report and cross validation score, we choose Random Forest Classifier as the best model for our dataset. Let’s see what were the results for Random Forest Classifier.

#RANDOM FOREST CLASSIFIER

# Checking accuracy for Random Forest Classifier

RFC = RandomForestClassifier ()

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

# Prediction

predRFC = RFC.predict(x\_test)

print (accuracy\_score(y\_test, predRFC))

print (confusion\_matrix(y\_test, predRFC))

print(classification\_report(y\_test,predRFC))

0.939568345323741

[[335 15]

[ 27 318]]

precision recall f1-score support

0 0.93 0.96 0.94 350

1 0.95 0.92 0.94 345

accuracy 0.94 695

macro avg 0.94 0.94 0.94 695

weighted avg 0.94 0.94 0.94 695

In [ ]:

The Accuracy score for Random Forest Classifier is 94%.

In [91]:

# Let’s plot confusion matrix for RandomForestClassifier

cm = confusion\_matrix(y\_test,predRFC)

x\_axis\_labels = ["0","1"]

y\_axis\_labels = ["1","0"]

f, ax = plt. subplots(figsize= (7,7))

sns.heatmap(cm, annot = True, linewidths=.2, linecolor="black", fmt = ".0f", ax=ax, cmap="BuGn",xticklabels=x\_axis\_labels,yticklabels=y\_axis\_labels)

plt. xlabel("PREDICTED LABEL")

plt.ylabel("TRUE LABEL")

plt.title('Confusion Matrix for RandomForestClassifier')

plt.show()

A green and white squares with numbers

Description automatically generated

After selecting our classifier model, we went ahead with hyper parameter tuning.

**5.2 Hyperparameter Tuning**

* Usually, we only have a vague idea of the best hyperparameters and thus the best approach to narrow our search is to evaluate a wide range of values for each hyperparameter. Using Scikit-Learn’s GridSearchCV method, we can define a grid of hyperparameter ranges, and randomly sample from the grid.

GridSearchCV(cv=5, estimator=RandomForestClassifier(),

param\_grid={'criterion': ['gini'], 'max\_depth': [2, 4, 6],

'n\_estimators': [100]})

**In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.   
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.**

GridSearchCV

GridSearchCV (cv=5, estimator=RandomForestClassifier(),

param\_grid={'criterion': ['gini'], 'max\_depth': [2, 4, 6],

'n\_estimators': [100]})

estimator: RandomForestClassifier

RandomForestClassifier()

RandomForestClassifier

RandomForestClassifier()

In [106]:

GCV.best\_params\_

{'criterion': 'gini', 'max\_depth': 6, 'n\_estimators': 100}

In [107]:

from sklearn import metrics

from sklearn.metrics import mean\_absolute\_error

from sklearn.metrics import mean\_squared\_error

from sklearn.metrics import r2\_score

In [108]:

Attrition = RandomForestClassifier (criterion='gini', max\_depth=6, n\_estimators=100)

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

pred = Attrition.predict(x\_test)

print (accuracy\_score(y\_test, predRFC))

print ("RMSE value:",np.sqrt(metrics.mean\_squared\_error(y\_test,pred)))

print('R2\_Score:',r2\_score(y\_test,pred)\*100)

0.939568345323741

RMSE value: 0.3263047963850438

R2\_Score: 57.407867494824025

The accuracy level is 94%, not much difference.

**5.2 Plotting AUC and ROC curve for the models.**

AUC — ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve, and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. The green line represents the ROC curve of a purely random classifier; a good classifier stays as far away from that line as possible (toward the top-left corner).

A graph of a number of different colored lines

Description automatically generated

As shown above, the fine-tuned **Random Forest Classifier** model showed a higher AUC score compared to other Algorithms.

**6.Concluding Remarks**

As the company generates more data on its employees (on New Joiners and recent Leavers) the algorithm can be re-trained using the additional data and theoretically generate more accurate predictions to identify **high-risk employees** of leaving based on the probabilistic label assigned to each feature variable by the algorithm.

**Indicators and Strategic Retention Plan**

The stronger indicators of people leaving include:

* **Monthly Income**: people on higher wages are less likely to leave the company. Hence, efforts should be made to gather information on industry benchmarks in the current local market to determine if the company is providing competitive wages.
* **Over Time**: people who work overtime are more likely to leave the company. Hence efforts must be taken to appropriately scope projects upfront with adequate support and manpower to reduce the use of overtime.
* **Age**: Employees in relatively young age bracket 25–35 are more likely to leave. Hence, efforts should be made to clearly articulate the long-term vision of the company and young employees fit in that vision, as well as provide incentives in the form of clear paths to promotion for instance.
* **Total Working Years**: The more experienced employees are less likely to leave. Employees who have between 5–8 years of experience should be identified as potentially having a higher-risk of leaving.
* **Years At Company**: Employees who hit their two-year anniversary should be identified as potentially having a higher risk of leaving.
* **YearsWithCurrManager**: Many leavers leave 6 months after being with their Current Managers. By using Line Manager details for each employee, one can determine which Manager have experienced the largest numbers of employees resigning over the past year.
* **Work Life Balance**: May female employees tend to have issue with work life balance, they can use some sort of counselling so that they could work productively at the same time, they are also projecting higher risk of leaving.
* **Tailored Retention Programs**: Customizing retention programs based on the specific needs of different employee groups (e.g., high performers, long-tenure employees, and employees in high-risk roles) can help reduce overall turnover rates.
* **Regular Monitoring**: Organizations should continue to monitor employee data regularly to identify emerging attrition trends and address them in a timely manner.
* **Cross-functional Collaboration**: HR, management, and department heads should collaborate to implement the necessary changes based on the analysis insights to enhance employee satisfaction and retention.

In conclusion, HR attrition analysis not only improves employee retention but also helps build a more resilient, engaged, and productive workforce. By leveraging data insights, organizations can create a sustainable employee retention strategy that aligns with their long-term goals.