Human Resources Analytics

Avetti Commerce

By – Akshay Gulhane

Human Resources Analytics

❖ An Introduction

❖ ***Definition –***

➢ *HR analytics is the process of collecting and analyzing Human Resource (HR) data in order to improve an organization's workforce performance. The process can also be referred to as talent analytics, people analytics, or even workforce analytics. This method of data analysis takes data that is routinely collected by HR and correlates it to HR and organizational objectives. Doing so provides measured evidence of how HR initiatives are contributing to the organization's goals and strategies.*

❖ ***Need of HR Analytics –***

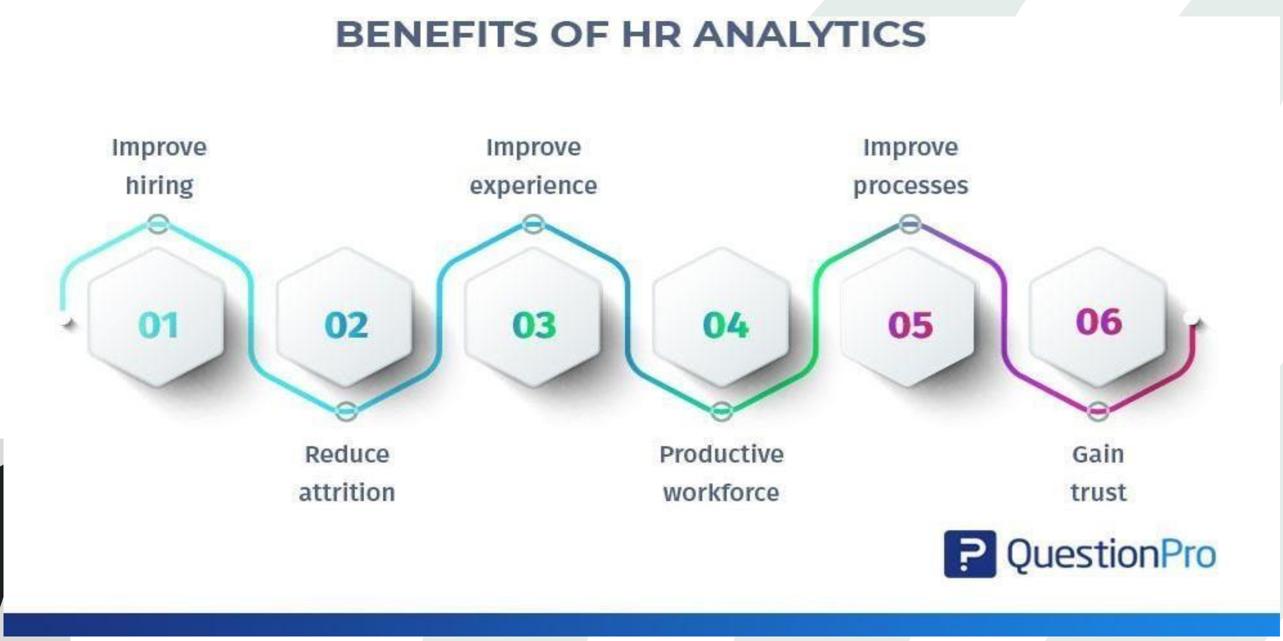
➢ *Most organizations already have data that is routinely collected, so why the need for a specialized form of analytics? Can HR not simply look at the data they already have? Unfortunately, raw data on its own cannot actually provide any useful insight. It would be like looking at a large spreadsheet full of numbers and words. Once organized, compared and analyzed, this raw data provides useful insights. They can help answer questions like*:

- **What patterns can be revealed in employee turnover?**
- **How long does it take to hire employees?**
- **What amount of investment is needed to get employees up to a fully productive speed?**
- **Which of our employees are most likely to leave within the year?**
- **Are learning and development initiatives having an impact on employee performance?**

❖ ***The process of HR Analytics*** -

➢ *HR Analytics is made up of several components that feed into each other.*

- **To gain the problem-solving insights that HR Analytics promises, data must first be collected.**
- **The data then needs to be monitored and measured against other data, such as historical information, norms or averages.**
- **This helps identify trends or patterns. It is at this point that the results can be analyzed at the analytical stage.**
- **The final step is to apply insight to organizational decisions.**



Employee Attrition

❖ An Introduction

- ❖ *Employee attrition analytics is specifically focused on identifying why employees voluntarily leave, what might have prevented them from leaving, and how we can use data to predict attrition risk. Most importantly, this type of employee predictive analytics can be used to help organizations understand and design the interventions that will be most effective in reducing unwanted attrition.*

- ❖ *In this project, **I am going to Analyze & Predict Employee Attrition**, whether an employee will leave the organization or will stay in the organization & I will present my Findings, Insights, Final conclusions & Business Impact of having this model from Analysis, Predictions & Model Evaluation in this presentation. I am using IBM HR Analytics Dataset from Kaggle for this project.*

❖ *About Task -*

➢ *We have to uncover the **factors that lead to employee attrition** and we have to **explore important questions regarding attrition of employees** by doing Exploratory Data Analysis, Data Visualization & based on it by we have classify between the employees who are leaving the organization & employees who are staying in the organization by using appropriate Machine Learning Model after Evaluating the Model based on **Accuracy Score & Area Under the ROC Curve**.*

❖ *About Dataset*

➢ *This is a fictional data set created by IBM data scientists.*

- **We have the following columns in our Dataset –**

Education
1 'Below College'
2 'College'
3 'Bachelor'

4 'Master'
5 'Doctor'

EnvironmentSatisfaction
1 'Low'
2 'Medium'
3 'High'
4 'Very High'

JobInvolvement
1 'Low'
2 'Medium'
3 'High'
4 'Very High'

JobSatisfaction
1 'Low'
2 'Medium'
3 'High'
4 'Very High’

PerformanceRating
1 'Low'
2 'Good'
3 'Excellent'
4 'Outstanding'

RelationshipSatisfaction
1 'Low'
2 'Medium'
3 'High'
4 'Very High'

WorkLifeBalance
1 'Bad'
2 'Good'
3 'Better'
4 'Best'

- ***There are 1470 Rows & 35 Columns in our Dataset***

Findings & Insights from Analysis & Visualization

Insights from Summary Statistics of the Data -

- *The average age of employees at IBM is 39, which means while hiring, they prefer candidates with decent work experience and expect higher level of expertise.*
- *The average salary hike for employees is 15% with maximum being 25%. With decent salary hike in the organization, employees tend to stay longer at the company and tend to enjoy long-term benefits with job security. This means, IBM rewards it's employees for their performance. This is proportional to employee satisfaction.*
- *However, the average Employee satisfaction stands at 2.7 out of 5.*
- *Most of the employees who get into IBM have worked with 2 or 3 companies in the past.*
- *On an average, an employee has worked at IBM for around 11 years and there seems to be an outlier - wherein an employee has worked for 38 years.*
- *It takes around 2 years for an IBM employee to bag his/her next promotion at the workplace.*

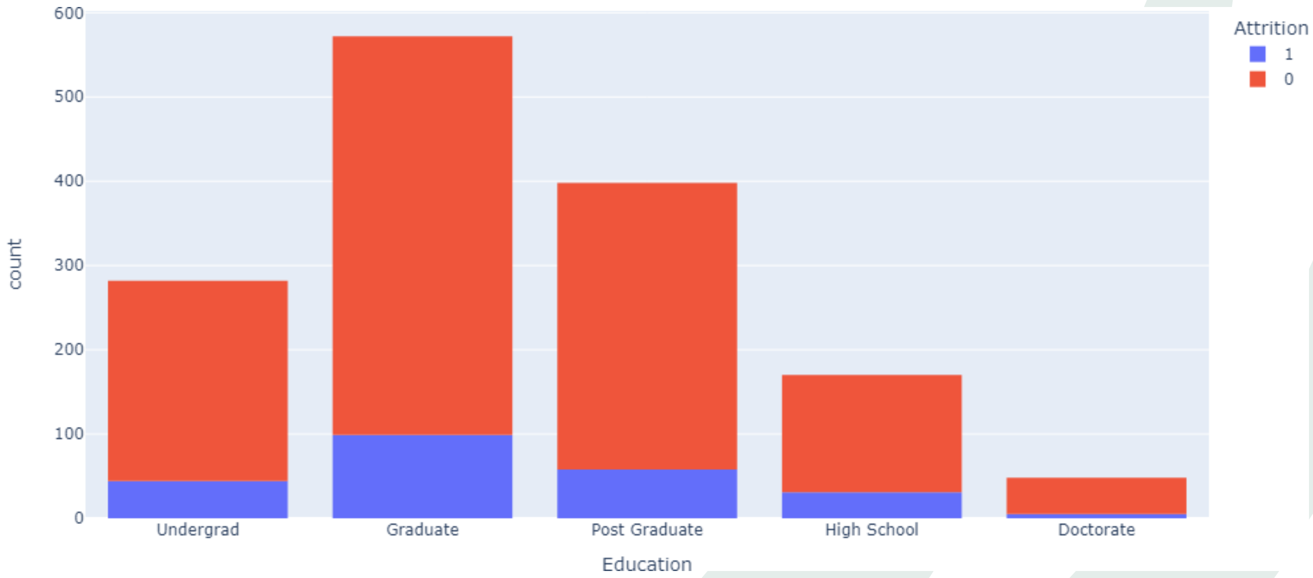
Insights from Correlation Matrix -

- *Age and monthly income are highly correlated.*
- *Age and number of years of Experience are highly correlated.*
- *Income is highly correlated with working hours*
- *Salary hike is highly correlated with Performance Bonus*

Insights from comparison between the employees who stayed & who left the Organization –

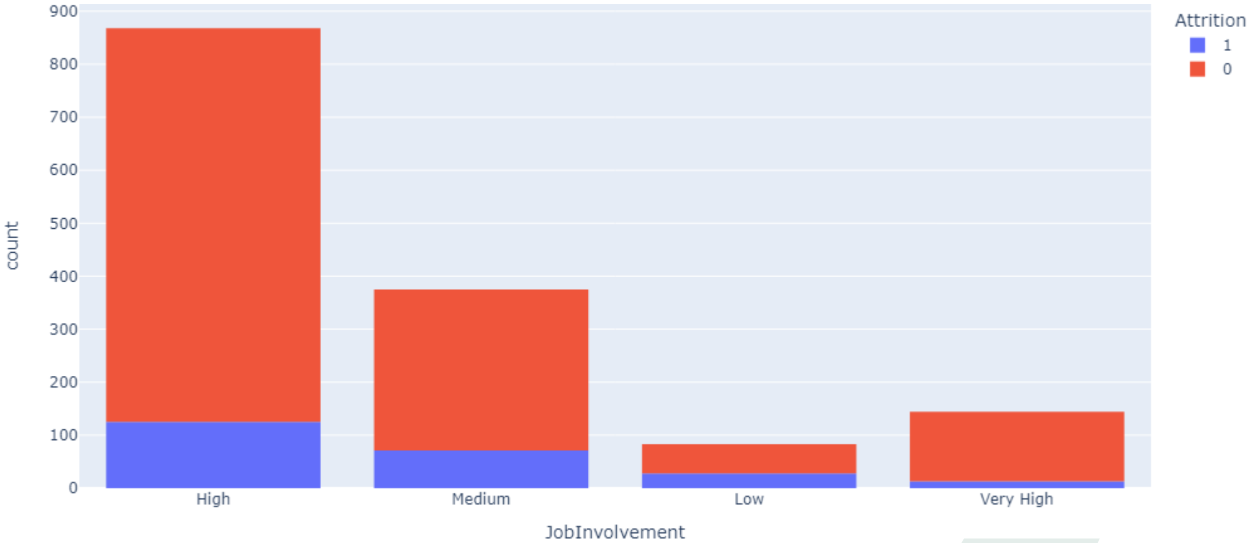
- *Mean age of the employees who stayed is higher compared to wholeft*
- *DailyRate of employees who stayed is higher*
- *'DistanceFromHome': Employees who stayed live closer to home*
- *'EnvironmentSatisfaction' & 'JobSatisfaction': Employees who stayedare generally more satisifed with their jobs*
- *'StockOptionLevel': Employees who stayed tend to have higher stockoption level*

❖ *Attrition by Education Level –*



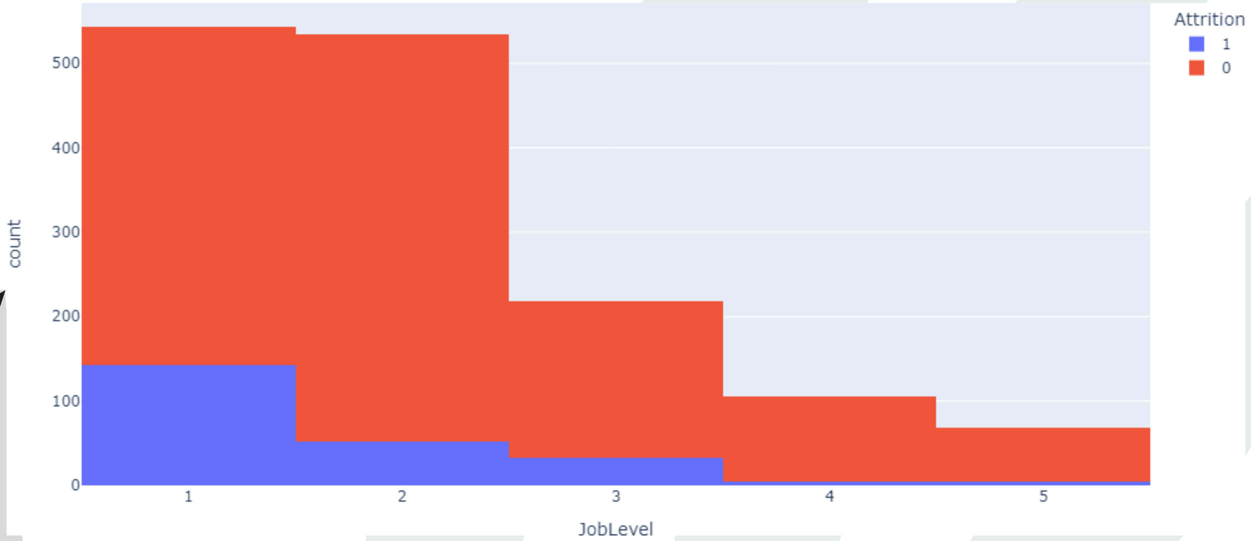
- ***Finding** – Mostly Graduates Employees seems to be leaving the company.*

❖ *Attrition by JobInvolvement –*



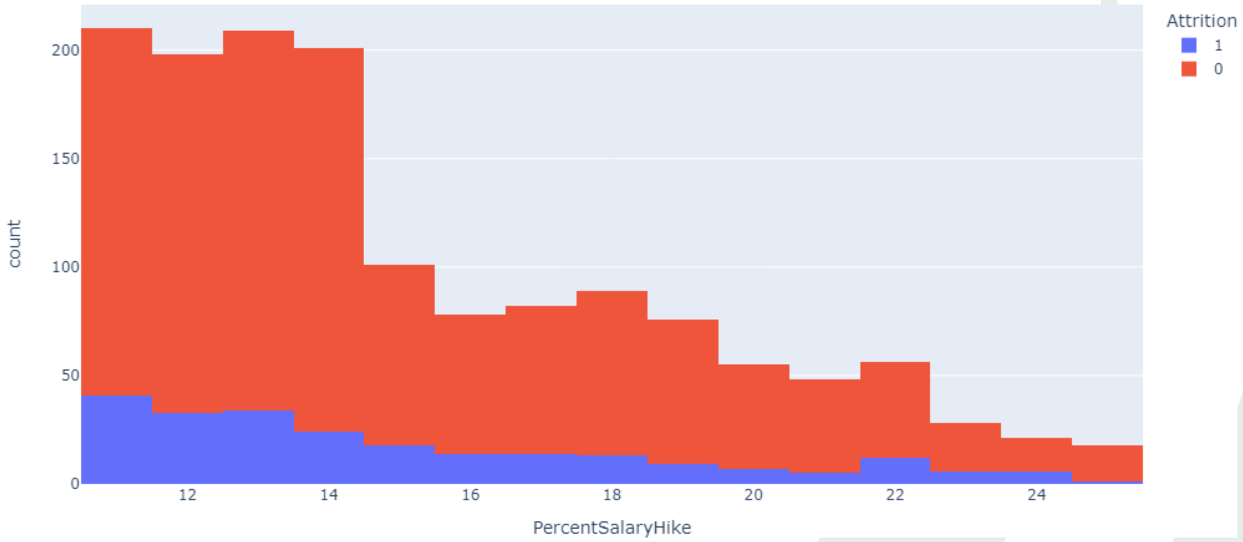
➢ ***Finding*** - *Employees with High Job Involvement tend to leave the company, while there is less attrition between Low and very high level job involvement.*

❖ *Attrition by Job Level–*



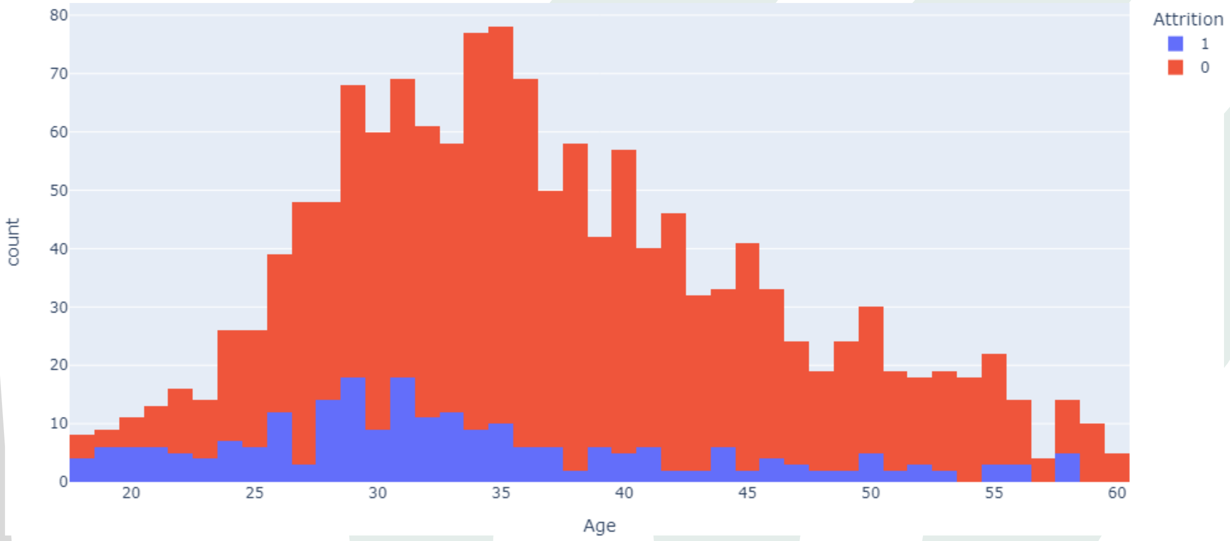
- ***Finding*** - *Mostly Employees between junior and associate level tend to leave the company more than compared to others.*

❖ *Attrition by Salary Hike –*



- ***Finding*** - *Employees who receive hikes between 12-14% tend to leave the company more than others with higher Salary hike percentage.*

❖ *Attrition by Age –*



- ***Finding*** - *Majority of the employees who leave the company are less than 40yrs of age.*

❖ *Attrition by Specific Job Role –*

- ***Finding*** - *As there is less salary to Sales Representative, the attrition for this job role is at higher side.*

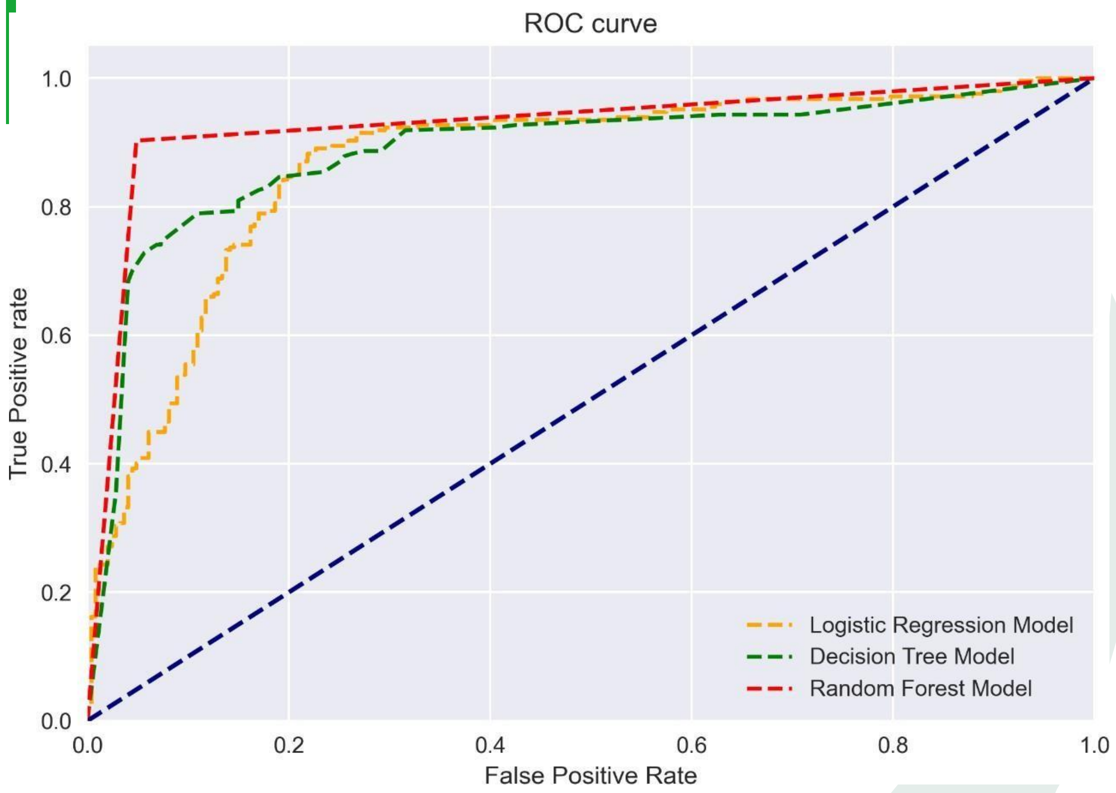
Final Conclusions

❖ *Reason Behind Choosing Random Forest Classifier Algorithm*

❖ ***"Random Forest Classifier Model" has the Highest Accuracy Score of 92% & AUC Score of 92% , which indicates that it has the Highest Accuracy & HighestArea Under the Curve and is the Best Model at Correctly Classifying Observations into Categories among Decision Tree Classifier Model & Logistic Regression Model.***

❖ ***I have explained about the terms Accuracy Classification Score, Confusion Matrix & 'AUC' (the area under 'ROC' curve) below for Model Evaluation for Selecting the Best Performing Model for our Employee Attrition ClassificationTask.***

❖ ***Our "Random Forest Classifier Model" has been Evaluated by these Performance Metrics & It is the Best Performing Model as it is Giving HighestAccuracy Score & Highest AUC RUC Score among Decision Tree Classifier & Logistic Regression Models after passing all Performance Metrics which we have selected for Model Evaluation.***



❖ *Accuracy Classification Score –*

- *A classification report is a performance evaluation metric in machine learning. It is used to show the precision, recall, F1 Score, Accuracy and support of your trained classification **model**.*
- *It provides a better understanding of the overall performance of our trained model.*
- ***Model accuracy** is a machine learning classification model performance metric that is defined as the ratio of true positives and true negatives to all positive and negative observations. In other words, accuracy tells us how often we can expect our machine learning model will correctly predict an outcome out of the total number of times it made predictions. For example: Let's assume that you were testing your machine learning model with a dataset of 100 records and that your machine learning model predicted all 90 of those instances correctly. The accuracy metric, in this case, would be: (90/100) = 90%. The accuracy rate is great but it doesn't tell us anything about the errors our machine learning models make on*

new data we haven't seen before.

- *Mathematically, it represents the ratio of the sum of true positive and true negatives out of all the predictions.*
- ***Accuracy Score = (TP + TN)/ (TP + FN + TN + FP)***

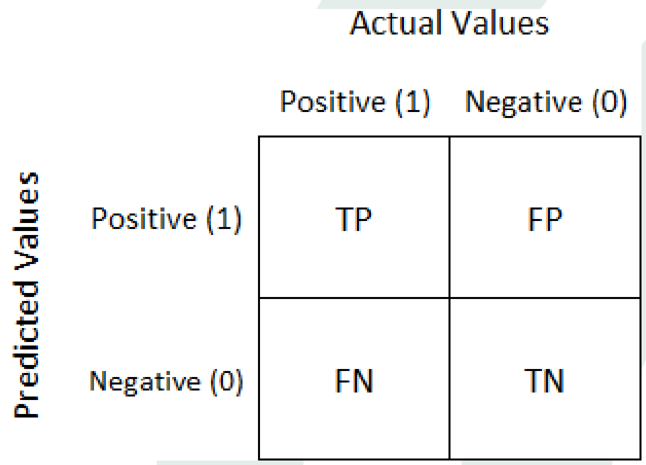
- ***Accuracy** is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. **Yes, accuracy is a great measure but only when you have symmetric datasets** where values of false positive and false negatives are almost same.*

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

❖ *Confusion Matrix representing Predictions vs Actuals –*

What is Confusion Matrix?

- Well, it is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual value.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

- **It is extremely useful for measuring Recall, Precision, Specificity, Accuracy, and most importantly AUC-ROC curves.**

➢ ***True Positive (TP):***

- **Interpretation: You predicted positive and it's true.**

- **You predicted that a woman is pregnant and she actually is.**

- True positive measures the extent to which the model correctly predicts the positive class. That is, the model predicts that the instance is positive, and the instance is actually positive. True positives are relevant when we want to know how many positives our model correctly predicts. For example, in a binary classification problem with classes "A" and "B", if our goal is to predict class "A" correctly, then a true positive would be the number of instances of class "A" that our model correctly predicted as class "A". Taking a real-world example, if the model is designed to predict whether an email is spam or not, a true positive would occur when the model correctly predicts that an email is a spam. The true positive rate is the percentage of all instances that are correctly classified as belonging to a certain class. True positives are important because they indicate how well our model performs on positive instances.

➢ ***False Positive (FP):***

- **Interpretation: You predicted positive and it's false.**
- **You predicted that a man is pregnant but he actually is not.**
- False positives occur when the model predicts that an instance belongs to a class that it actually does not. False positives can be problematic because they can lead to incorrect decision-making. For example, if a medical diagnosis model has a high false positive rate, it may result in patients undergoing unnecessary treatment. False positives can be detrimental to classification models because they lower the overall accuracy of the model.

➢ ***True Negative (TN):***

- **Interpretation: You predicted negative and it's true.**
- **You predicted that a man is not pregnant and he actually is not.**
- True negatives are the outcomes that the model correctly predicts as negative.

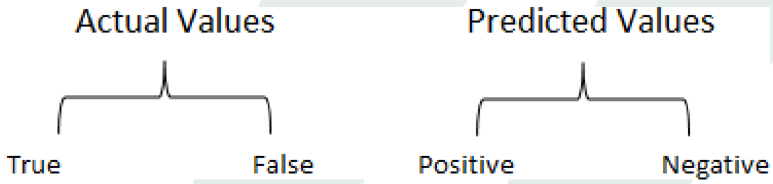
For example, if the model is predicting whether or not a person has a disease, a true negative would be when the model predicts that the person does not have the disease and they actually don't have the disease. True negatives are one of the measures used to assess how well a classification model is performing. In general, a high number of true negatives indicates that the model is performing well. True negative is used in conjunction with false negative, true positive, and false positive to compute a variety of performance metrics such as accuracy, precision, recall, and F1 score. While true negative provides valuable insight into the classification model's performance, it should be interpreted in the context of other metrics to get a complete picture of the model's accuracy.

- ***False Negative (FN):***

- **Interpretation: You predicted negative and it's false.**
- **You predicted that a woman is not pregnant but she actually is.**

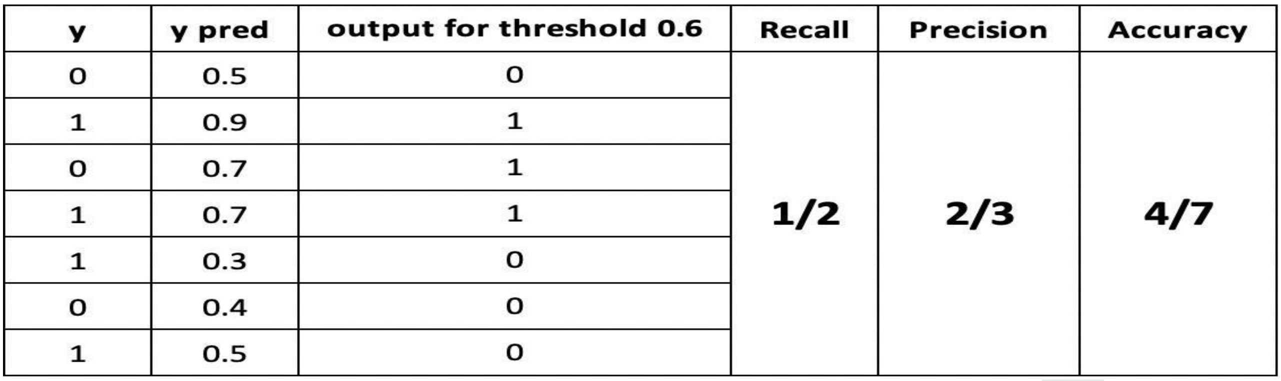
A false negative occurs when a model predicts an instance as negative when it is actually positive. False negatives can be very costly, especially in the field of medicine. For example, if a cancer screening test predicts that a patient does not have cancer when they actually do, this could lead to the disease progressing without treatment. False negatives can also occur in other fields, such as security or fraud detection. In these cases, a false negative may result in someone being granted access or approving a transaction that should not have been allowed. False negatives are often more serious than false positives, and so it is importantto take them into account when evaluating the performance of a classification model.

- **Just Remember, We describe predicted values as Positive and Negative and actual values as True and False.**



- ***How to Calculate Confusion Matrix for a 2-class classification problem?***

- Let’s understand the confusion matrix through math.



y	y pred	output for threshold 0.6	Recall	Precision	Accuracy
0	0.5	0	1/2	2/3	4/7
1	0.9	1			
0	0.7	1			
1	0.7	1			
1	0.3	0			
0	0.4	0			
1	0.5	0			

- **Recall**

$$\textit{Recall} = \frac{TP}{TP + FN}$$

- The above equation can be explained by saying, from all the positive classes, how many we predicted correctly.

- Recall should be high as possible.

- **Precision**

$$\textit{Precision} = \frac{TP}{TP + FP}$$

- The above equation can be explained by saying, from all the classes we have predicted as positive, how many are actually positive.

- Precision should be high as possible.

- **Accuracy**

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

- From all the classes (positive and negative), how many of them we have predicted correctly.

- Accuracy should be high as possible.

- **F-measure**

$$\textit{F - measure} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall + Precision}}$$

- *It is difficult to compare two models with low precision and high recall or vice versa. So to make them comparable, we use F-Score. F-score helps to measure Recall and Precision at the same time. It uses Harmonic Mean in place of Arithmetic Mean by punishing the extreme values more.*

❖ *Caution with Accuracy Metrics / Score* -

- *The following are some of the* ***issues with accuracy metrics*** */ score:*
- **The same accuracy metrics for two different models may indicate different model performance towards different classes.**
- **In case of imbalanced dataset, accuracy metrics is not the most effective metrics to be used.**
- One should be **cautious when relying on the accuracy metrics** of model to evaluate the model performance.

- *The **accuracy metrics is also not reliable** for the models trained on **imbalanced datasets.** Take a scenario of dataset with 95% imbalance (95% data is negative class). The accuracy of the classifier will be very high as it will be correctly doing right prediction issuing negative most of the time. A better classifier that actually deals with the class imbalance issue, is likely to have a worse accuracy metrics score. In such scenario of **imbalanced dataset**, another metrics **AUC (the area under ROC curve) is more robust than the accuracy metrics** score. The AUC takes into the consideration, the class distribution in imbalanced dataset.*

❖ *'AUC' (the area under 'ROC' curve) –*

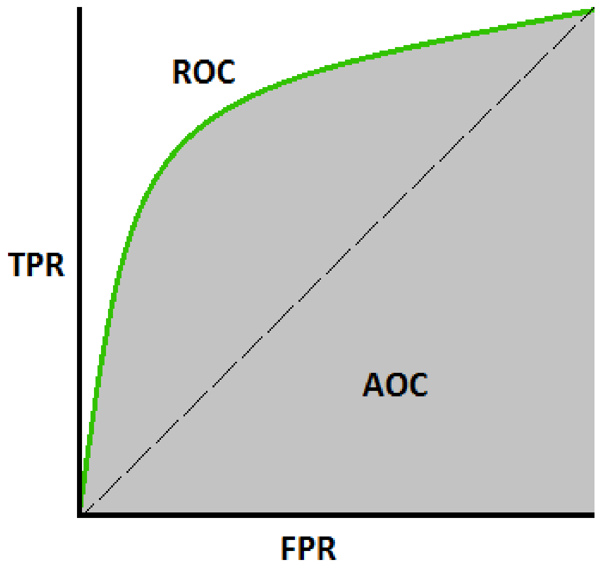
- ***Understanding AUC - ROC Curve***

- ***In Machine Learning, performance measurement is an essential task. So when it comes to a classification problem, we can count on an AUC - ROC Curve. When we need to check or visualize the performance of the multi-class classification problem, we use the AUC (**Area Under the Curve**) ROC (**Receiver Operating Characteristics**) curve. It is one of the most important evaluation metrics for checking any classification model's performance. It is also written as AUROC (**Area Under the Receiver Operating Characteristics**)***

- ***What is the AUC - ROC Curve?***

- AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. By analogy, the Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease.

- The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis.



➢ *Defining terms used in AUC and ROC Curve.*

- **TPR (True Positive Rate) / Recall /Sensitivity**

$$\text{TPR /Recall / Sensitivity} = \frac{TP}{TP + FN}$$

- **Specificity**

$$\text{Specificity} = \frac{TN}{TN + FP}$$

- **FPR (False Positive Rate)**

$$\text{FPR} = 1 - \text{Specificity}$$

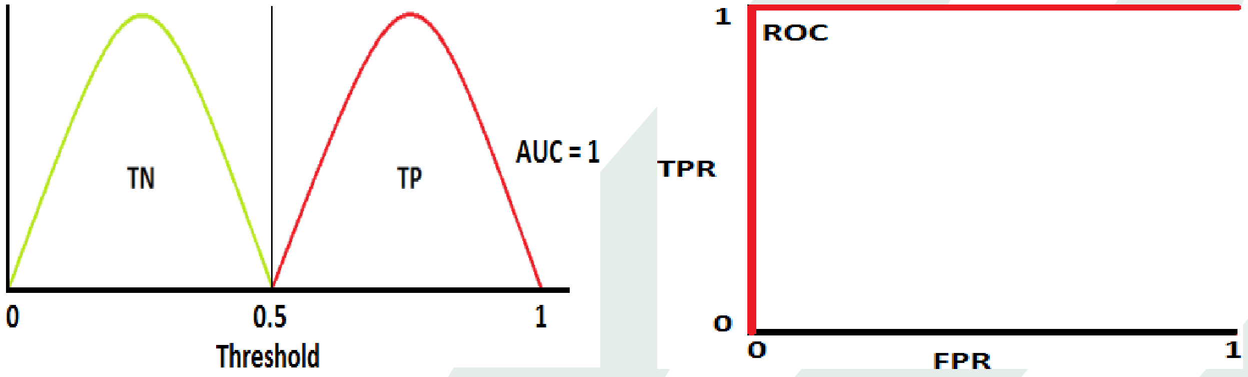
$$= \frac{FP}{TN + FP}$$

➢ *How to speculate about the performance of the model?*

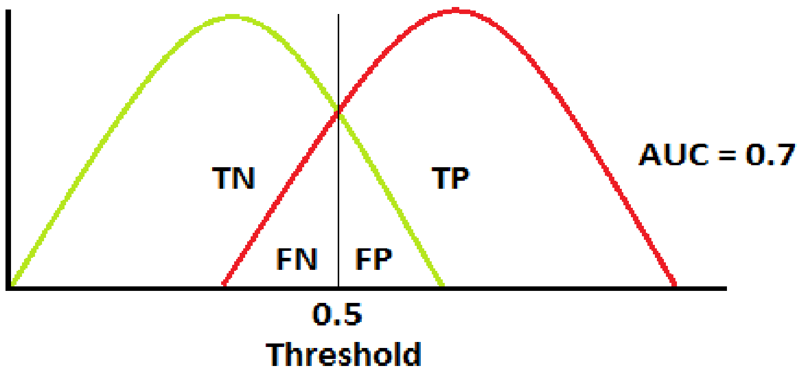
- An excellent model has AUC near to the 1 which means it has a good measure of separability. A poor model has an AUC near 0 which means it has the worst measure of separability. In fact, it means it is reciprocating the result. It is predicting 0s as 1s and 1s as 0s. And when AUC is 0.5, it means the model has no class separation capacity whatsoever.

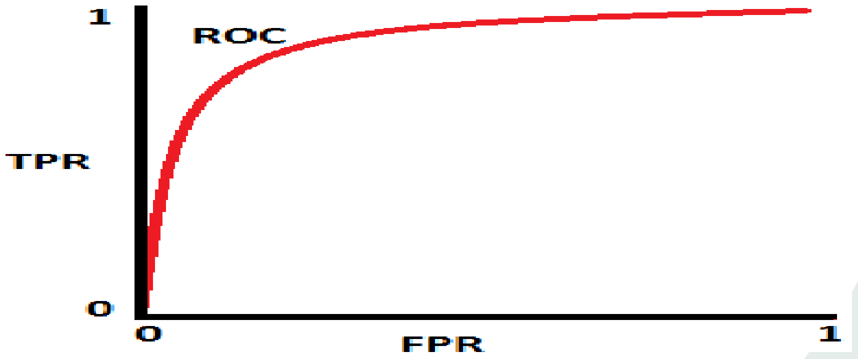
- Let's interpret the above statements.
- As we know, ROC is a curve of probability. So let's plot the distributions of those probabilities:

- Note: Red distribution curve is of the positive class and the green distribution curve is of the negative class.

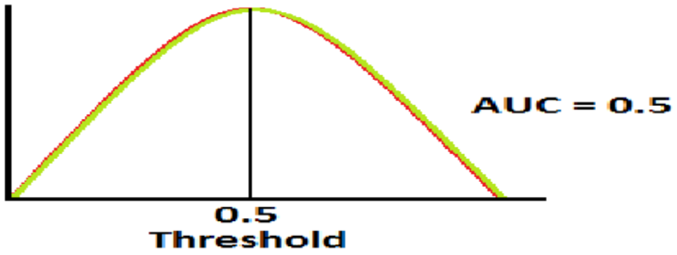


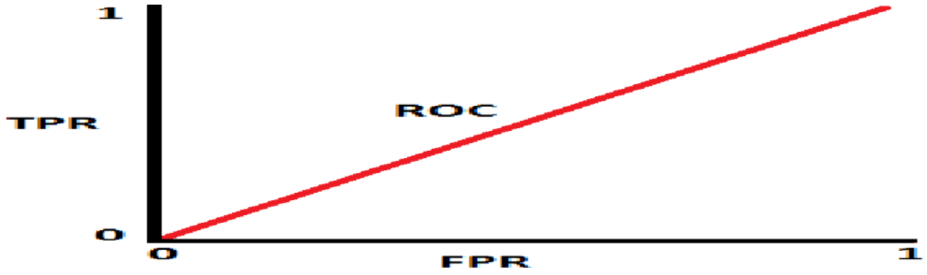
- This is an ideal situation. When two curves don't overlap at all means model has an ideal measure of separability. It is perfectly able to distinguish between positive class and negative class.



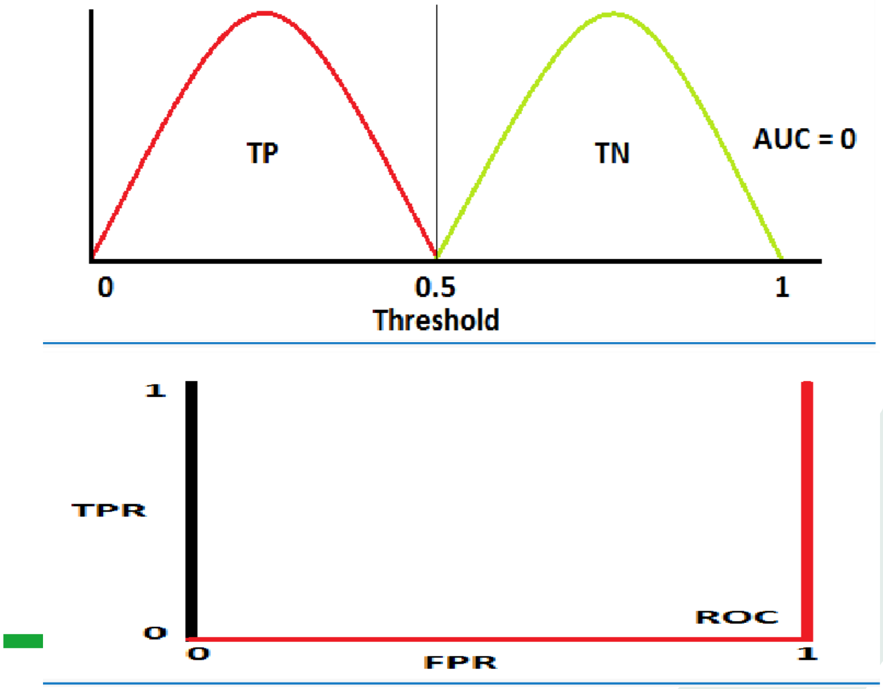


- When two distributions overlap, we introduce type 1 and type 2 errors. Depending upon the threshold, we can minimize or maximize them. When AUC is 0.7, it means there is a 70% chance that the model will be able to distinguish between positive class and negative class.





- This is the worst situation. When AUC is approximately 0.5, the model has no discrimination capacity to distinguish between positive class and negative class.



- When AUC is approximately 0, the model is actually reciprocating the classes. It means the model is predicting a negative class as a positive class and vice versa.

➢ ***The relation between Sensitivity, Specificity, FPR, and Threshold.***

- Sensitivity and Specificity are inversely proportional to each other. So when we increase Sensitivity, Specificity decreases, and vice versa.

- SensitivityY, Specificityt and Sensitivityt, SpecificityY

- When we decrease the threshold, we get more positive values thus it increases the sensitivity and decreasing the specificity.

- Similarly, when we increase the threshold, we get more negative values thus we get higher specificity and lower sensitivity.

- As we know FPR is 1 - specificity. So when we increase TPR, FPR also increases and vice versa.

- TPRY, FPRY and TPRt, FPRt

Business Impact of Having this Model Handy for a Tech Firm

❖ *Attrition in an Organization | Why Workers Quit?*

➢ *Employees are the backbone of the organization. Organization's performance is heavily based on the quality of the employees. Challenges that an organization has to face due employee attrition are:*

- **Expensive in terms of both money and time to train new employees.**
- **Loss of experienced employees**
- **Impact in productivity**
- **Impact profit**

❖ *Business questions to Brainstorm -*

➢ ***What factors are contributing more to employee attrition?***

➢ ***What type of measures should the company take in order to retain their employees?***

➢ ***What business value does the model bring?***

- *Will the model save lots of money?*
- *Which business unit faces the attrition problem?*

-

❖ *Impact of Having this Final Random Forest Model Handy for a Tech Firm*

- ***Accuracy Score of Random Forest Model is 85% and it is the Best Model at Correctly Classifying Observations into Categories whether an Employee will Leave the Organization or will Stay in the Organization, among Decision Tree & Logistic Regression Models.***

- ***It will be very Usefull & Impactfull Model for Predicting Employee Attrition.***

- ***As we have done exploratory data analysis, we have already found answers to above questions. We have already tuned our model according to our employee attrition problem & business solution requirement. We have taken into consideration only those features in our model which are causing the attrition of employees for better prediction of attrition.***

- ***Based on the above findings, insights & predictions from model, company should take appropriate measures & make a business decision in order to retain their employees.***

- ***This model will bring immense business value to the company as it will predict in advance as who will leave the company with great accuracy so that company can take the preventive measures in advance.***

- ***So by taking appropriate measures, it will reduce the cost for the company depending upon how the company can make use of the predictions of the model.***

- ***Sales representative unit or sales department facing the most attrition problem. So by taking appropriate measures company can solve the employee attrition problem.***

- ***By this way & by using <u>Technology for Problem Solving</u> for our <u>Employee Attrition Problem</u>, the Model will bring <u>Valuable Predictive Analytical Output</u> for the tech firm, by using the Output of the Model & by making <u>Preventive Business Decisions based on the Output of the Model</u> it will be very <u>Impactfull for the Business</u> as it will <u>Save theTime and Future Cost of the Firm</u> & the Output of the Model will be very <u>Valuable for the Firm.</u>***