**Airline Passenger Satisfaction Prediction**

**Aswini Kumar Motpothula,**

**Data science trainee,**

**Edureka Learning Center , Hyderabad**

**Abstract:**

**Airline Passenger Satisfaction Prediction: A Data-Driven Approach**

**Airline passenger satisfaction is a critical aspect of the aviation industry, directly impacting customer loyalty, brand reputation, and overall profitability. Predicting passenger satisfaction is a complex task due to the multifaceted nature of the airline experience, involving factors such as flight punctuality, in-flight services, customer service, and baggage handling**.

***Keywords: Airline passenger satisfaction, Machine learning, Predictive modeling, Feature engineering***

**1.Problem Statement**

**Airline passenger satisfaction is a crucial factor in the aviation industry, impacting customer loyalty and overall business performance. However, accurately predicting passenger satisfaction levels based on various factors is a complex task. The problem addressed in this research is the need to develop a robust and accurate predictive model using machine learning techniques that can anticipate and quantify passenger satisfaction.**

**Existing research has highlighted the significance of factors such as flight punctuality, in-flight services, customer service, and baggage handling in determining passenger satisfaction. However, there is a lack of comprehensive predictive models that effectively capture the interplay of these factors and provide accurate predictions**

**Additionally, with the increasing volume of customer reviews, survey responses, and operational data collected by airlines, there is a need for advanced techniques to extract meaningful insights from the available data and develop predictive models that can handle the complexity and variability of passenger satisfaction.**

## **2.Data Description**

### Attribute Information:

### **Here is a list of potential attributes that could be included in the dataset for predicting airline passenger satisfaction::**

* **Gender:** Gender of the passengers (Female, Male)
* **Customer Type**: The customer type (Loyal customer, disloyal customer)
* **Age:** The actual age of the passengers
* **Type of Travel:** Purpose of the flight of the passengers (Personal Travel, Business Travel)
* **Class:** Travel class in the plane of the passengers (Business, Eco, Eco Plus)
* **Flight distance:** The flight distance of this journey
* **Inflight wifi service:** Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)
* **Departure/Arrival time convenient:** Satisfaction level of Departure/Arrival time convenient
* **Ease of Online booking:** Satisfaction level of online booking
* **Gate location**: Satisfaction level of Gate location
* **Food and drink**: Satisfaction level of Food and drink
* **Online boarding**: Satisfaction level of online boarding
* **Seat comfort**: Satisfaction level of Seat comfort
* **Inflight entertainment:** Satisfaction level of inflight entertainment
* **On-board service:** Satisfaction level of On-board service
* **Leg room service:** Satisfaction level of Leg room service
* **Baggage handling:** Satisfaction level of baggage handling
* **Check-in service:** Satisfaction level of Check-in service
* **Inflight service:** Satisfaction level of inflight service
* **Cleanliness:** Satisfaction level of Cleanliness
* **Departure Delay in Minutes:** Minutes delayed when departure
* **Arrival Delay in Minutes:** Minutes delayed when Arrival
* **Satisfaction:** Airline satisfaction level(Satisfaction, neutral or dissatisfaction)

**3. Introduction**

**Airline passenger satisfaction is a critical aspect of the aviation industry, as it directly impacts customer loyalty, brand reputation, and overall profitability. Satisfied passengers are more likely to become repeat customers and recommend the airline to others, while dissatisfied passengers may switch to competitors and share their negative experiences, affecting the airline's bottom line. Therefore, accurately predicting and understanding passenger satisfaction levels is of utmost importance for airlines to enhance their services and improve the overall passenger experience.**

## **4. Reasons for dissatisfactions**

**The reasons for dissatisfactions are:**

* **Poor Customer Service**
* **Poor Cleanliness and Hygiene**
* **Overall Service Quality**

**5. Most Dissatisfies Age people**

**1. most of the 78th age passengers dissatisfies**

**2.least dissatisfies were above 80age passengers**

**6. Steps involved:**

* **Exploratory Data Analysis**

**After loading the dataset we performed this method by comparing** **our target** **variable that is Sastisfaction with Other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable**

* **Null values Treatment**
* **Our dataset contains Arrival Delay variable contain null values filled with Median**
* **Encoding of categorical columns**

**We used One Hot Encoding to produce binary integers of 0 and 1 to encode our categorical features are Gender, Customer Type, Type of Travel,Class,Satisfaction, because categorical features that are in string format cannot be understood by the machine and**

* **Feature Selection**

**In these steps we using information gaining method finding correlation coefficients between variables mostly effects that are removed from our data set i.e**  **" Arrival Delay " removed from the dataset**

* **APPLYING SMOTE**

**The given data set unbalanced data set for this we used SMOTE i.e** Synthetic Minority Oversampling Technique **resampling of to performing best results for randomly before after applying SMOTE**

**Original dataset shape Counter({0: 58762, 1: 45142})**

**Resample dataset shape Counter({0: 73452, 1: 73452})**

**Counter({0: 73452, 1: 73452})**

**AND fitting and applying different algorithms to it**

**The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.**

* **Fitting different models**

**For modeling we tried various classification algorithms like:**

1. **Logistic Regression**
2. **KNN Classifier**
3. **Random Forest Classifier**
4. **Decision Tree**
5. **Gaussian NB**

* **Tuning the hyperparameters for better accuracy**

**Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting in case of tree based models**

**like Random Forest Classifier and XGBoost classifier**.

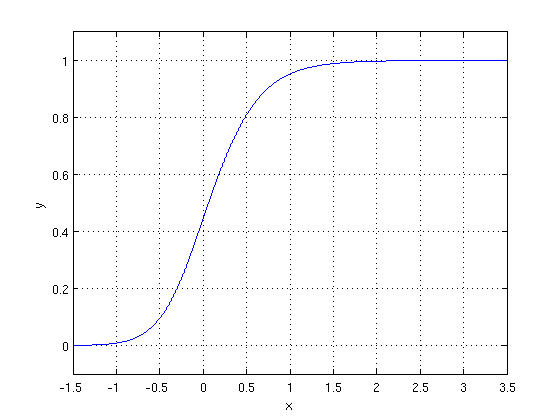
**7.1. Algorithms:**

1. **Logistic Regression:**

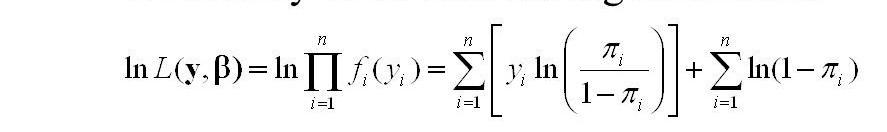
**Logistic Regression is actually a classification algorithm that was given the name regression due to the fact that the mathematical formulation is very similar to linear regression.**

**The function used in Logistic Regression is sigmoid function or the logistic function given by:**

**f(x)= 1/1+e ^(-x)**



**The optimization algorithm used is: Maximum Log Likelihood. We mostly take log likelihood in Logistic:**



**We have implemented logistic regression with Grid search cv. We get an accuracy score of approximately 62%. and precision score approximately is 62% and f1\_score is 62%and roc auc approximately is 62% As we have an imbalanced dataset, recall score is approximately 63% better parameter.**

.

1. **KNN Classifier:**

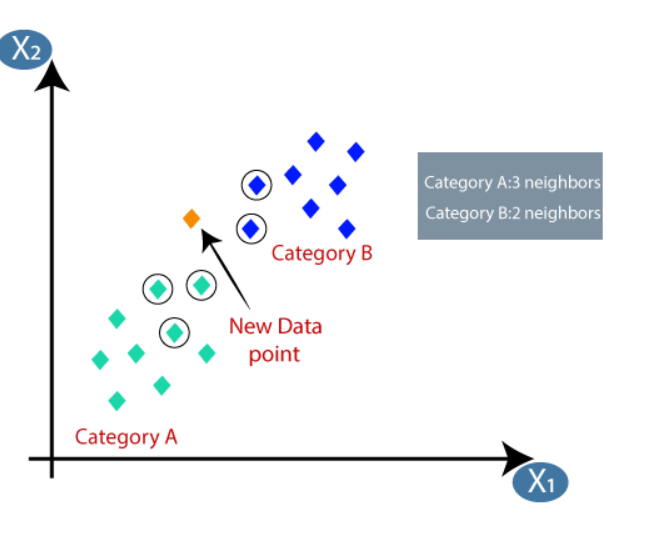
* **K-Nearest Neighbors is one of the simplest supervised learning algorithms.**

### **The KNN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories.**

### **KNN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suited category by using KNN algorithm.**

### **KNN algorithms can be used for regression as well as for classification problems.**

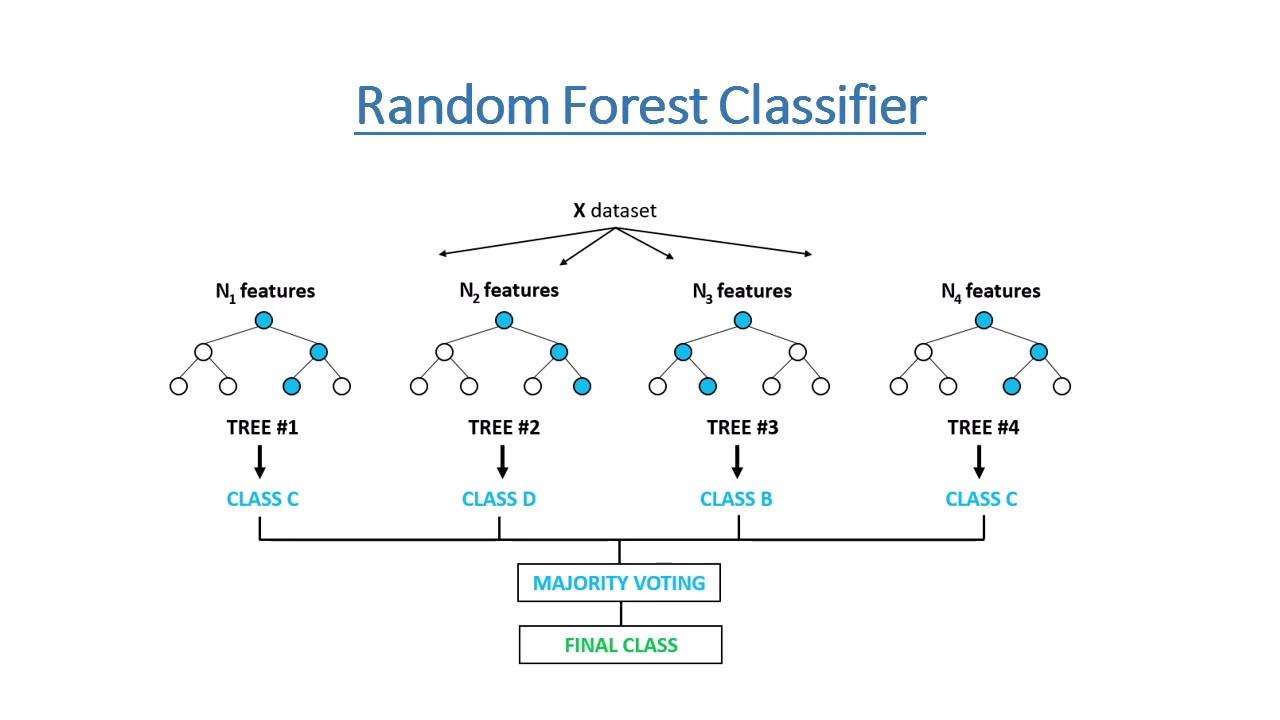
### **As we can see the 3 nearest neighbors are from category A, hence this new data point must belong to category A**



**KNN and our accuracy score is approximately 75%. and precision score is approximately 70% and f1\_score is 76% and ROC\_AUC score is 75% ,recall score is approximately 88% better parameter**

1. **Random Forest Classifier:**

**Random Forest is a bagging type of Decision Tree Algorithm that creates a number of decision trees from a randomly selected subset of the** **training set,** **collects the labels from these subsets and then averages the final prediction depending on the most number of times a label has been predicted out of all.**



**Random Forest and our accuracy score is approximately 86%. and recall score is approximately 82% and f1\_score is 85% and ROC\_AUC score is 86% ,precision score is approximately90% better parameter. Let's go ahead with other models and see if they can give better result.**

**number of leaves T in each tree (so that in the above** example, **T=3 and w=[2, 0.1, -1]).**

**When building a decision tree, a challenge is to decide how to split a current leaf. For instance, in the above image, how could I add another layer to the (age > 15) leaf? A ‘greedy’ way to do this is to consider every possible split on the remaining features (gender and occupation), and calculate the new loss for each split; you could then pick the tree which most reduces your loss.**

**7.2. Model performance:**

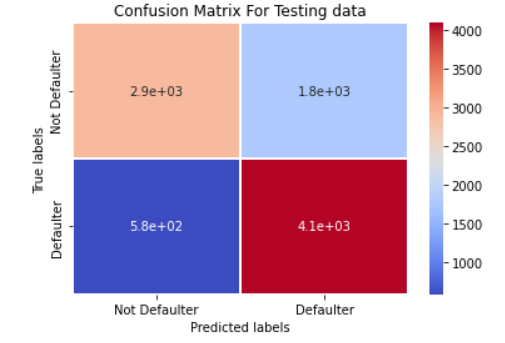
**Model can be evaluated by various metrics such as:**

1. **Confusion Matrix**-

**The confusion matrix is a table that summarizes how successful the classification model is at predicting examples belonging to various classes. One axis of the confusion matrix is the label that the model predicted, and the other axis is the actual label.**

**Out of all models Random forest classifier with Grid search CV performed very well its confusion matrix of test data is**





1. **Precision/Recall**-

**Precision is the ratio of correct positive predictions to the overall number of positive predictions : TP/TP+FP**

**Recall is the ratio of correct positive predictions to the overall number of positive examples in the set: TP/FN+TP**

**Out of them Random forest classifier with Grid search CV ( tuned hyperparameters gave) the best result. it's Highest**

**Precision score is approximately 90%,**

**Recall score is approximately 82%**

1. **Accuracy**-

**Accuracy is given by the number of correctly classified examples divided by the total number**

**of classified examples. In terms of the confusion matrix, it is given by: TP+TN/TP+TN+FP+FN**

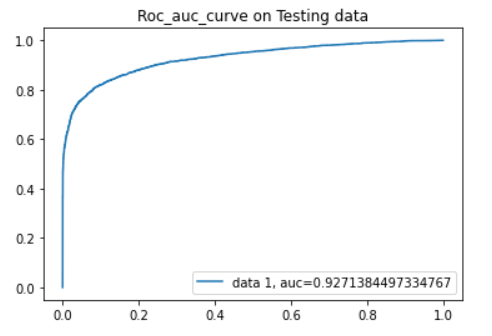
**Out of themRandom forest classifier Grid search CV ( tuned hyperparameters gave) the best result. it's Highest**

**Accuracy Score approximately 86%,**

1. **Area under ROC Curve(AUC)**-

**ROC curves use a combination of the true positive rate (the proportion of positive examples predicted correctly, defined exactly as recall) and false positive rate (the proportion of negative examples predicted incorrectly) to build up a summary picture of the classification performance.**

**Out of them Random forest classifier with Grid search CV ( tuned hyperparameters gave) the best result. it's Highest** **ROC\_AUC score is approximately 86%,**

****

**7.3. Hyper parameter tuning:**

**Hyperparameters are sets of information that are used to control the way of learning an algorithm. Their definitions impact parameters of the models, seen as a way of learning, change from the new hyperparameters. This set of values affects performance, stability and interpretation of a model. Each algorithm requires a specific** **hyperparameters grid that can be adjusted according to the business problem. Hyperparameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.**

**We used Grid Search CV, for hyperparameter tuning. This also results in cross validation and in our case we divided the dataset into different folds. The best performance improvement .**

1. **Grid Search CV-Grid Search combines a selection of hyperparameters established by the scientist and runs through all of them to evaluate the model’s performance. Its advantage is that it is a simple technique that will go through all the programmed combinations. The biggest disadvantage is that it traverses a specific region of the parameter space and cannot understand which movement or which region of the space is important to optimize the model.**

**8. Conclusion:**

**That's it! We reached the end of our exercise.**

**Starting with loading the data so far we have done EDA , null values treatment, encoding of categorical columns, feature selection and then model building.**

**In all of these models our accuracy revolves in the range of 62 to 96%.**

**And there is no such improvement in accuracy score even after hyperparameter tuning.**

**We used different type of Classification algorithms to train our model like, Logistic Regression, Random Forest Classifier, KNN Classifier, Classifier. and Also we tuned the parameters of Random forest classifier and XG boost classifier ,KNN\_ Out of them Random forest classifier with Grid search CV ( tuned hyperparameters gave) the best result. it's Highest**

**Precision score is approximately 90%,**

**Recall score is approximately82%**

**ROC\_AUC score is approximately 86%,**

**Accuracy Score is approximately 86%,**

**and It's F1\_score approximately 85%**