Airline passenger referral prediction

Zeeshan Ahmad
Data science trainee,
AlmaBetter, Bangalore

Abstract:

Air business as we know has been largely affected due to Covid-19 and most of the airline now is sitting on the verge of Bankruptcy because of this situation. Due to industrialization and modernization is a huge competition between airlines to retain their customers by providing best service. A customer review dataset consisting of around 17 features is given by almabetter to predict whether a customer will refer airline to his friend or not with the help of machine learning classification models.

1. Problem Statement

Data contains airline reviews from 2006 to 2019 from popular airlines around the world with multiple choices and free text questions. Data is scraped in spring 2019. The main objective is whether the passengers will refer the airline to their friends

2. Introduction

The Airline passenger Referral system has become the most important criteria globally for the airline industry in order to address the surge which has been created after global pandemic so as to remain in the global market competition. Airline referral system generally works on customer reviews which is basically sentiment given by the customer depending upon various factor like seat comfort, their trip distance, route they have travelled, timing, the airline frequency, ground service etc. on the basis of which sentiment reviews are analyzed and machine learning model on classification is prepared which helps airline industries to focus on the factor resolving which it can actually help them in business growth better than the competitors.

3. Data descriptions

airline: Name of the airline

overall: Overall point is given to the trip between 1 to 10

author: Author of the trip #reviewdate: Date of the Review customer.

review: Review of the customers in free text format.

aircraft: Type of the aircraft.

Traveller type: Type of traveler (e.g. business, leisure)

cabin: Cabin at the flight.

date flown: Flight date

seat comfort: Rated between 1-5

cabin service: Rated between 1-5.

foodbev: Rated between 1-5

entertainment: Rated between 1-5

ground service: Rated between 1-5 value for money: Rated between 1-5.

recommended: Binary, target variable.

4. Steps involved:

Data collection:

Data collection is the process of collecting, measuring and analyzing different types of information using a set of standard validated techniques. The main objective of data collection is to gather information-rich and reliable data, and analyses them to make critical business decisions. Once the data is collected, it goes through a rigorous process of data cleaning and data processing to make this data truly useful for businesses. It refers to the process of finding and loading data into our system. Pandas library is used to loading our data in our system in python. Using pandas we can manipulate data easily.

Data Cleaning:

The next task was data cleaning which was easy with this dataset. Data cleaning refers to the process of removing unwanted variables and values from your dataset and getting rid of any irregularities in it. Such anomalies can disproportionately skew the data and hence adversely affect the results. Some steps that can be done to clean data are:

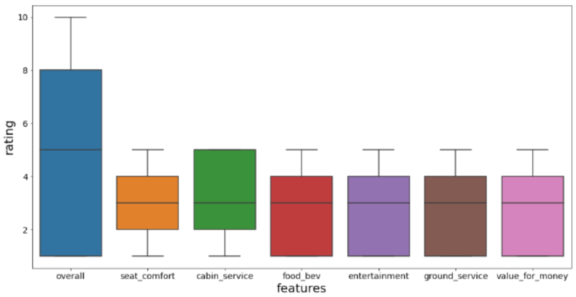
- Handling missing values: There are always some missing values in dataset. If we don’t remove or handle those missing values then that can cause a trouble in our analysis. Removing or replacing those missing values with something meaningful is very important so that our data will have no missing values.
- Removing duplicates: Drop the duplicates rows.
- Formatting data to proper dtype.
- Adding or removing columns required for analysis.

Exploratory Data Analysis:

Exploratory Data Analysis (EDA) plays a vital role in the analysis of the data variables which are important from the aspect of feature engineering. It will help us to distribute and relate between dependent and independent variables. We have gone through an analysis of every independent as well as the dependent variable to check which independent factor affects the dependent factor.

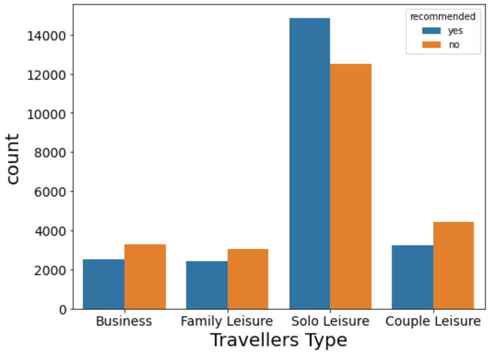
4.1. Outlier detection

Outlier detection is the process of detecting and subsequently excluding outliers from a given set of data. An outlier may be defined as a piece of data or observation that deviates drastically from the given norm or average of the data set. An outlier may be caused simply by chance, but it may also indicate measurement error or that the given data set has a heavy-tailed distribution.



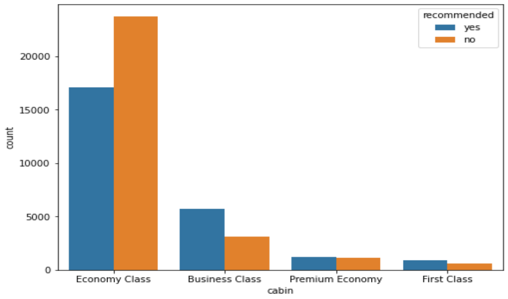
The outliers are not present in our data.

4.2. Travelers flight recommendation



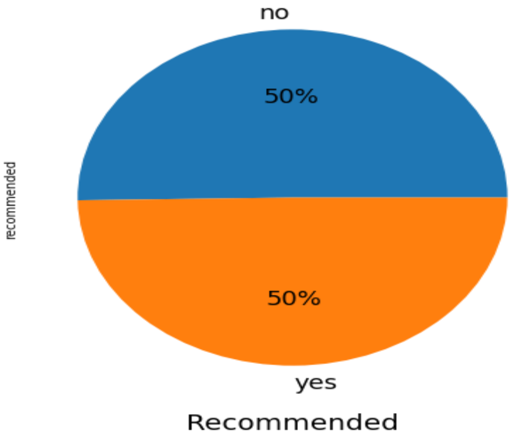
Travellers from solo leisure are recommended yes more than others travellers. Travellers from business and family leisure have less recommendation. Negative recommendation is also high from solo leisure.

4.3. Which type of   cabin has maximum recommendations?



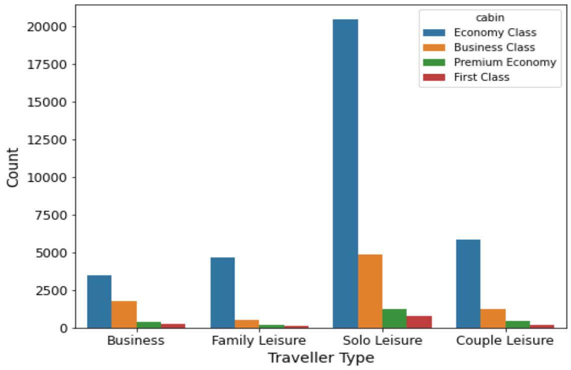
For the Economy class, Number of 'NO'   recommendations are more than 'YES' recommendations. For business class and first class, the Number of 'YES' recommendations are more than 'NO' recommendations. For the Premium account number of 'YES' recommendations and 'NO' recommendations are approximately equal.

4.4. Total recommendation percentage for all airlines?



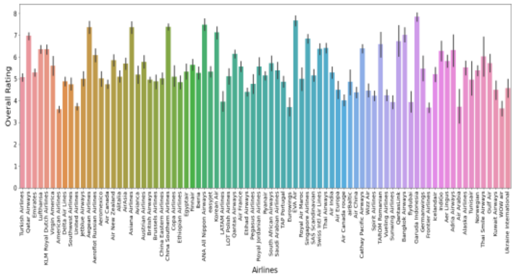
The overall recommendation percentage for all airlines is 50% which is equal to recommended 'NO'. The data is not imbalanced.

4.4. Travellers and their cabin



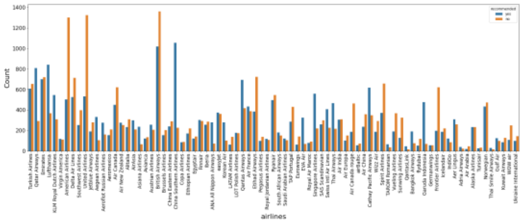
All types of travelers mostly prefer economy class. Business class is less preferrable than economy class. First class is least preferable among all travelers.

4.5. Overall rating given by different customers



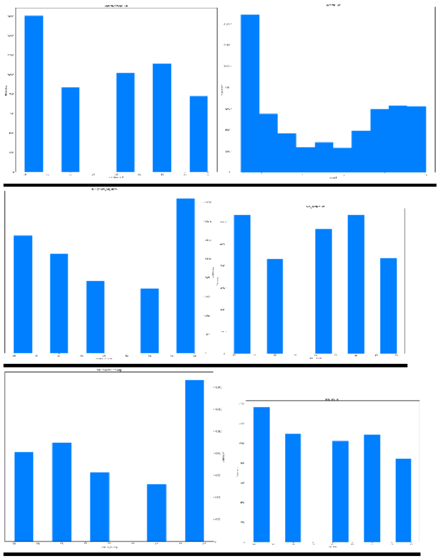
The maximum overall ratings are received by Aegean airlines, Asiana airlines, China southern Airlines, ANA ALL Nippon Airways, EVA Air, Garuda Indonesia (rating is around 7.5-8). The minimum overall rating is received by American airlines, United airlines, Eurowings, Frontier airlines, Air Arabia and WOW air.

4.6. Airlines and their recommendations



American airlines, United airlines and British airlines received maximum 'NO' recommendations. China southern airlines, Lufthansa, British airlines and Qatar airways received maximum 'YES' recommendations. Thai smile, Tunisair, Air Arabia, Adria airways received minimum 'Yes' recommendations.

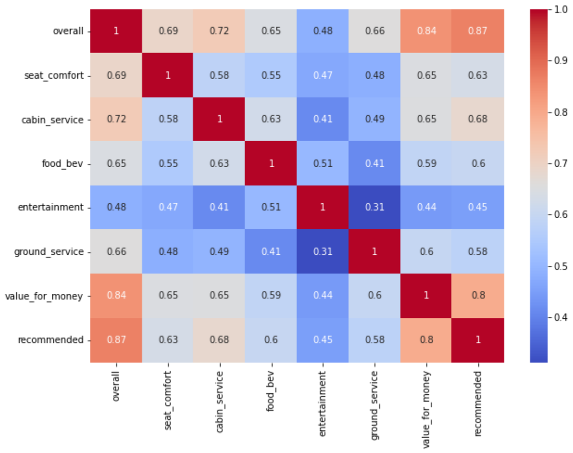
4.7. Frequency of values



1. Cabin service got the maximum rating of 5.
2. Overall rating got by the airlines is poor equal to 1.
3. Maximum customers rate food_bev as poor equal to 1.
4. Most of the customers have rated airlines as 1 indicating expensive(value for money).

4.7. Correlation

Let's check the heatmap plotted concerning independent variables.



'Overall’, 'food_bev’, 'cabin_service', 'value_for_money’, etc all are positively correlated with recommendation. 'Overall' is most correlated with recommendation. Entertainment has 0.45 of correlation which is minimum. Overall and value for money are multicollinear.

Model Training:

Model training is the process of fitting a data into machine learning model from which model learns the patterns in data to predict the dependent variable. Model do it so by assigning a weight to each variable. After our model is trained, we test our model on test data to check how our model is performing. 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. In this project we have used 80% data for training purpose and 20% data for test set. The train-test procedure is appropriate when there is a sufficiently large dataset available.

Fitting different models

For modeling we tried various regression algorithms like:

1. **Logistic Regression**
2. **Random forest**
3. **XG Boost Classifier**
4. **K-Nearest Neighbor**
5. **Random forest with GridSearch CV**
6. **KNN with GridSearch CV**

Tuning the hyperparameters for better accuracy

Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and avoiding overfitting. also called hyperparameter optimization, is the process of finding the configuration of hyperparameters that results in the best performance. The process is typically computationally expensive and manual.

We used Grid Search CV for hyperparameter tuning.

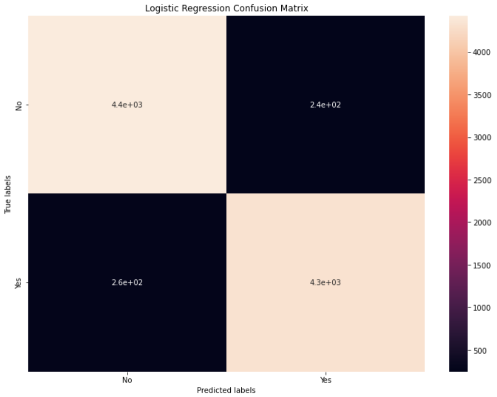
Grid Search CV:

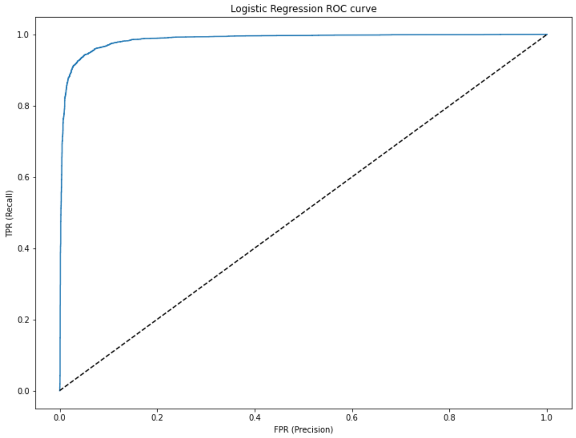
Grid Search combines a selection of hyperparameters established by the scientist and runs through all of them to evaluate the model’s performance. Its advantage is that it is a simple technique that will go through all the programmed combinations. The biggest disadvantage is that it traverses a specific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

5. Algorithms:

1. Logistic Regression:

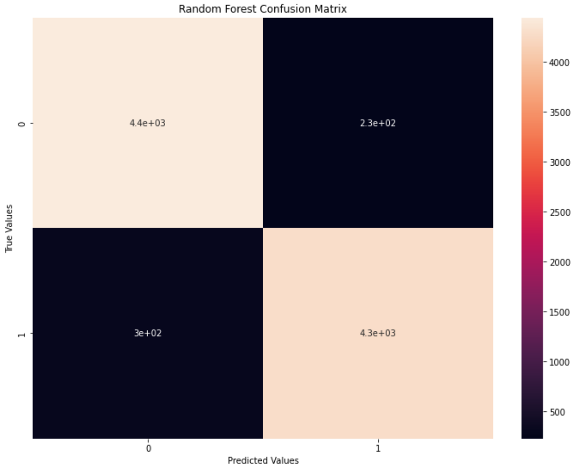
Logistic regression is a classification technique that predicts the likelihood of a single-valued result (i.e. a dichotomy). A logistic regression yields a logistic curve with values only ranging from 0 to 1. The likelihood that each input belongs to a specific category is modelled using logistic regression. Logistic regression is a fantastic tool to have in your toolbox for classification purposes. For classification situations, where the output value we want to predict only takes on a small number of discrete values, logistic regression is an important technique to know. The logistic function offers a number of appealing characteristics. The probability is represented by the y-value, which is always confined between 0 and 1, which is exactly what we wanted for probabilities. A 0.5 probability is obtained for an x value of 0. A higher likelihood is also associated with a higher positive x value, while a lower probability is associated with a greater negative x value. In logistic regression to learn the coefficients of features in order to maximize the probability of correctly classifying the classes. For this maximum likelihood, concept is used.

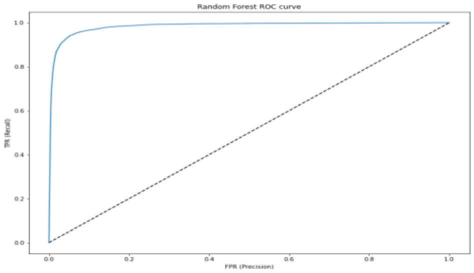




2. Random Forest:

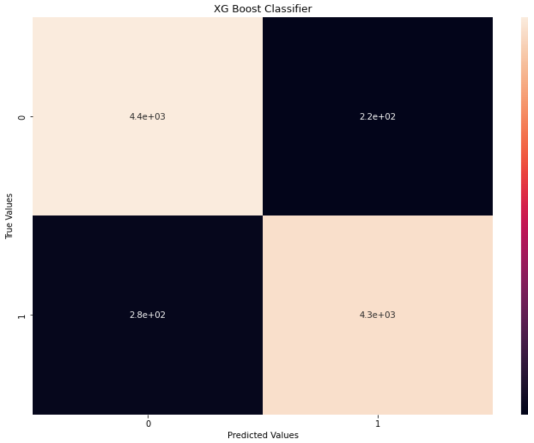
We create several trees in the Random Forest model rather than a single tree in the CART model. From the subsets of the original dataset, we create trees. These subsets can contain a small number of columns and rows. Each tree assigns a categorization to a new object based on attributes, and we say that the tree "votes" for that class. The classification with the highest votes is chosen by the forest.

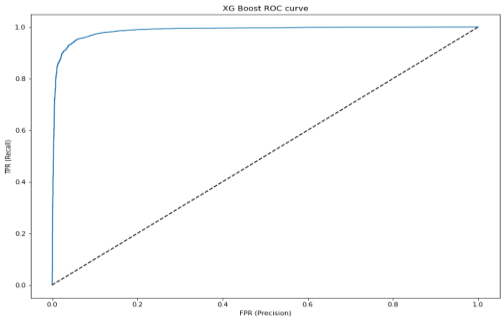




3. XG Boost Classifier

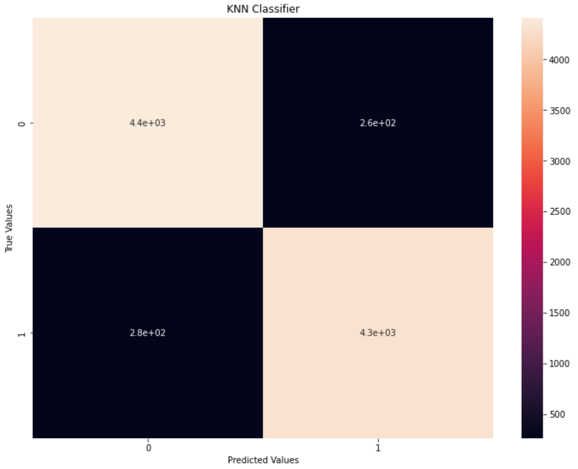
XG Boost is one of the most popular variants of gradient boosting. It is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. XG Boost is basically designed to enhance the performance and speed of a Machine Learning model. In prediction problems involving unstructured data (images, text, etc.), artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree-based algorithms are considered best-in-class right now.

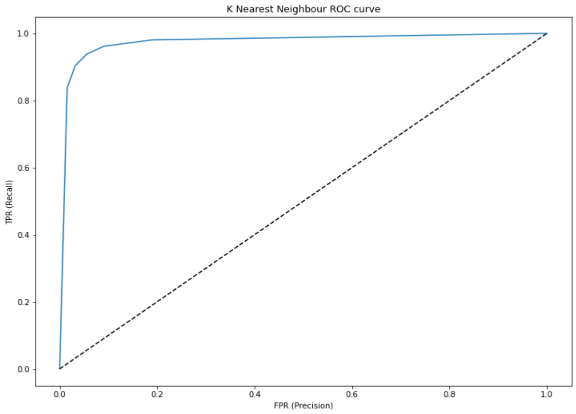




4. K_nearest Neighbour Model

K Nearest Neighbour is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure. It is mostly used to classifies a data point based on how its neighbours are classified.





6. Evaluation Metrics

Model can be evaluated by various metrics such as:

1. **Train Score:**
   A data with lots of variance then this causes over-fitting. This causes poor result on Test Score. Because the model curved a lot to fit the training data and generalized very poorly. So, generalization is the goal.

2. **Test Score:**
   This is when our model is ready. Before this step we have not touched this data-set. So, this represents real life scenario. Higher the score, better the model generalized.

3. **Accuracy:**
   Accuracy will require two inputs (i) actual class labels (ii)predicted class labels. To get the class labels from probabilities (these probabilities will be probabilities of getting a HIT), you can take a threshold of 0.5. Any probability above 0.5 will be labeled as class 1 and anything less than 0.5 will be labeled as class 0.

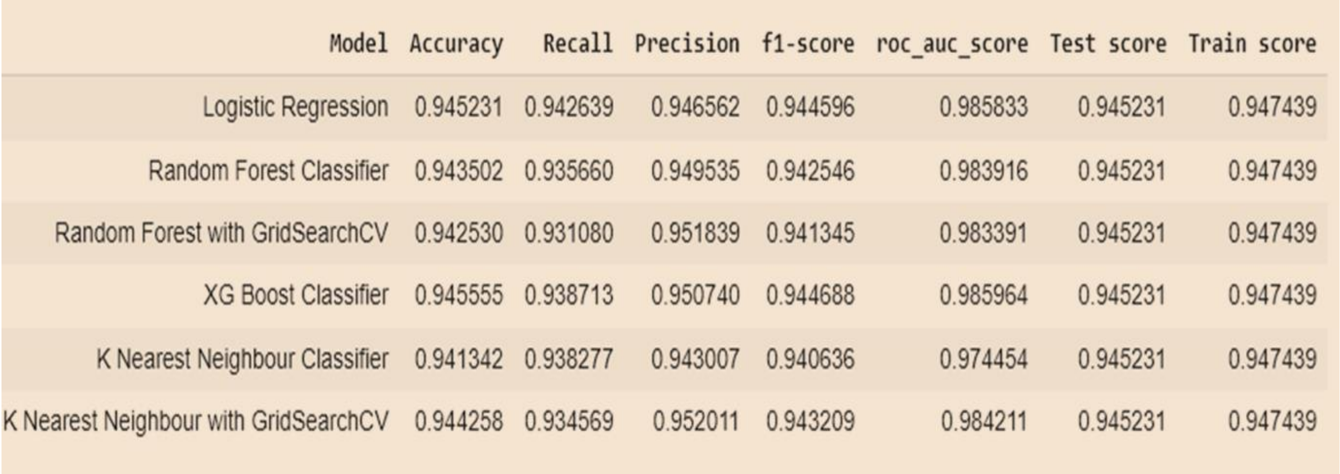
4. **Precision:**
   Precision for a label is defined as the number of true positives divided by the number of predicted positives. Report precision in percentages.
5. **Recall :**
   Recall for a label is defined as the number of true positives divided by the total number of actual positives. Report recall in percentages.

6. **F1-Score :**
   This is defined as the harmonic mean of precision and recall.

7. **AUC-ROC:**
   The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the ‘signal’ from the ‘noise’. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

7. Conclusion:

- 'Overall','foodbev', 'cabin_service', 'value_for_money', etc. are positively correlated with recommendation. These parameters should be improved to provide better service and hence it will improve recommendation chances for airlines.
- Entertainment has 0.45 of correlation which is less than others.
- 'Overall' is most correlated with recommendation.
- Multicollinearity is present in between overall and value_for_money.
- XG boost classifier gives better accuracy.
- XG Boost, Logistic Regression gives good results in terms of accuracy. The highest accuracy obtained is 0.9455 with XG Boost Classifier.
- KNN after applying hyperparameter tuning also gave good accuracy.
- American airlines, united airlines, spirit and British airlines received maximum 'NO' recommendations.
- China southern airlines, Lufthansa and Qatar airways received maximum 'YES' recommendations. Thai smile, Tunisair, Air Arabia, Adria airways received minimum 'Yes' recommendations.
- For Economy class, Number of 'NO' recommendations are more than 'YES' recommendations.
- For business class and first class, Number of 'YES' recommendations are more than 'NO' recommendations.
- For Premium economy class number of 'YES' recommendation and 'NO' recommendations are approximately equal.



Model	Accuracy	Recall	Precision	f1-score	roc_auc_score	Test score	Train score
Logistic Regression	0.945231	0.942639	0.946562	0.944596	0.985833	0.945231	0.947439
Random Forest Classifier	0.943502	0.935660	0.949535	0.942546	0.983916	0.945231	0.947439
Random Forest with GridSearchCV	0.942530	0.931080	0.951839	0.941345	0.983391	0.945231	0.947439
XG Boost Classifier	0.945555	0.938713	0.950740	0.944688	0.985964	0.945231	0.947439
K Nearest Neighbour Classifier	0.941342	0.938277	0.943007	0.940636	0.974454	0.945231	0.947439
K Nearest Neighbour with GridSearchCV	0.944258	0.934569	0.952011	0.943209	0.984211	0.945231	0.947439

References-

1. stackoverflow.com
2. towardsdatascience.com
3. GeeksforGeeks
4. Analytics Vidhya