Data Mining on Airline Reviews Dataset

Umaraj Potla
Computer Science
Georgia State University
Atlanta, Georgia
upotla1@student.gsu.edu

Sravya Kambhampati
Comupter Science
Georgia State University
Atlanta, Georgia
lkambhampati1@student.gsu.edu

Ruthvik Potu
Computer Science
Georgia State University
Atlanta, Georgia
rpotu1@student.gsu.edu

Taslim Murad
Computer Science
Georgia State University
Atlanta, Georgia
tmurad2@student.gsu.edu

***Abstract*—The objective of our study is to identify the important airline quality attributes from the online user ratings and review posts and to examine the effect of the airline quality attribute ratings on the airline recommendations. This study employed data-mining classification techniques and clustering on 27,000 passenger reviews to evaluate the service quality of airlines and whether the users recommend the airlines. Airline ratings dataset used for this study contains ratings on parameters like seat comfort, cabin staff, food quality, in-flight entertainment rating, value for money by passengers along with their personal reviews. Additionally, reviews and ratings data considered for this study include passengers from different cabin classes like Economy, Business Class, First class.**

I. Introduction

Airline ratings dataset was collected from Kaggle [1]. This dataset consists of more than 27 thousand customer ratings and reviews of 292 different airlines. More than 19 thousand user's reviews are recorded in this dataset with users from more than 190 countries. Airline's customers reviews were collected in years between 2013 to 2015. Data of quality ratings and reviews on airlines in the past have relied on subjective surveys of consumer opinion that were infrequently collected. Subjective approach in turn gives a quality rating that is essentially non-comparable for different surveys for any specific airline. The current study includes data from three different years with customers from all over the world and users travelling in different cabin classes of all airlines that reports actual airline performance on critical quality criteria important to consumers and combines them into a rating system. The result is a rating for individual airlines whether customers recommend using the airlines. Attributes of each airline that was considered for inclusion in the rating scale were basic criteria like seat-comfort, in-flight food and beverages quality, in-flight entertainment, cabin staff, value for money. Ratings in the dataset are on a scale of 0 to 5 in judging airline quality along with a generic overall-rating with a scale of 0 to 10. Ratings play an important role in consumer decision-making providing direction of impact that how strongly the consumer's rating of airline quality recommend the use of the particular airline. With the continued global trend in airline operations alliances, the study on Airline Rating can be used as a most dependent method of comparing the quality of airline performance for international operations as well and airlines can work on improving the particular criteria which received the lowest ratings.

II. Understanding The Dataset

A. Handling different types and scales of ratings

All ratings are not created on equal scales. Each criteria of Airline quality is rated on a scale of 0 to 5. Whereas, an attribute called overall rating of the airline varies on a scale of 0 to 10. Apart from the numeric values of ratings, customer ratings were from different cabin class flown. Ratings for each criteria of the airline might differ for differ cabin class. Cabin class attribute values are categorical like Economy, Business Class, Premium Economy and First Class. And since each rating source has a bias, it is required to normalize the scores so inputs are consistent.

Moreover to gain some general understanding of the data we made a few visualizations of each feature when put against the other features.

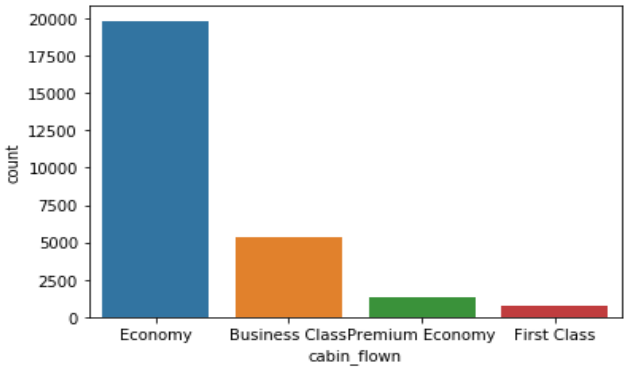

Fig. 1. Cabin Count Vs Cabin Flown

As per the plots in Fig 1 we can observe that the people have taken the Economy are the highest followed by Business Class, Premium Economy and First Class.

If we observe the plots in Fig 2 people who've recommended the particular airline have given a rating of 5 to the cabin staff where as who didn't have given a lower rating.

Based on the plots in Fig 3 people who have recommended an airline are more than people who have not.

According to the plots in Fig 4 out of the total count, people who have recommended a particular airline have rated food and beverages higher than people who have not.

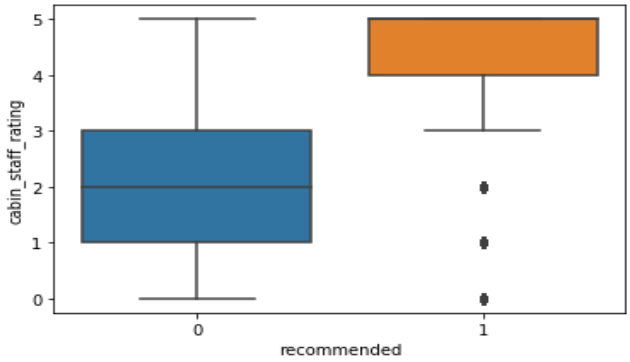

Fig. 2. Cabin Staff Rating Vs Recommended

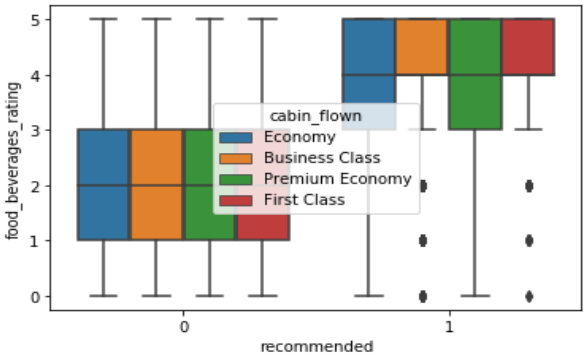

Fig. 4. Food and Beverages rating Vs Recommended

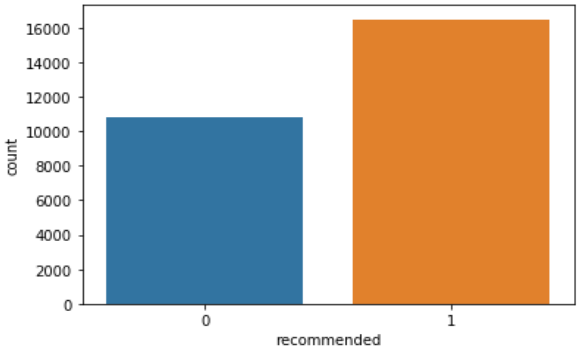

Fig. 3. Recommended count

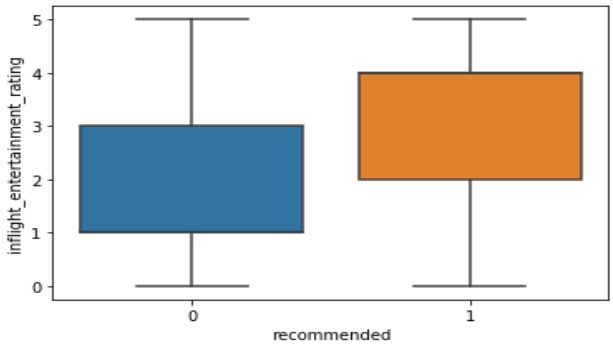

Fig. 5. Inflight Entertainment rating Vs Recommended

If we observe the plots in Fig 5 it's a mix in general but clearly we can see there's an edge to people recommending the airline when the rating is high.

If we observe the plots in Fig 6 people who have recommended the airline have rated the seat comfort higher compared to people who have not.

Based on the plots in Fig 7 we can clearly observe that people who have recommended an airline have given a high overall rating where as people who have not recommended have given a low overall rating.

Based on the plots in Fig 8 we can say that recommended and overall rating have the highest correlation.

III. Methodology

We applied the following classification models to predict if an airline is recommended or not: Decision Tree, Random Forests, Naive Bayes and Support Vector Machines. Along with classification, we also implemented clustering algorithms like PCA and K-Means. In this section we describe brief description about each model we chose followed by results and observations.

IV. Decision Trees Classification

A typical structure of the decision tree consists of a root node, multiple branches and number of leaf nodes. The dataset is incrementally divided into subsets until creating a decision tree with decision nodes and leaf nodes. Each node makes a decision as true or false question for one of the features of the dataset and the result to this decision divides the dataset into two subsets. Each internal node represents a test on the attribute and each branch represents the test result and each leaf node represents a class label. One of the challenges to build decision trees is to identify which feature to be selected as the next node for decision making. To overcome this challenge, decision tree classifier algorithm uses indices like entropy or Gini-impurity to quantify an uncertainty or impurity associated with a certain node.

A. Applying Decision Tree Classifier on Airline Dataset

As discussed in Section 1, airline ratings dataset with 27 thousand odd samples was used to train and test the three classifiers with the recommended - split. Decision Trees Classifier from sklearn tree library was used to train and test the model.

On applying Decision Tree Classifier algorithm on the dataset and building multiple models with 'Entropy' criterion and tree depths varying from 1 to 10, it is observed that the best DecisionTree Classifier model was built with a tree depth of 4 with most accuracy and where there is the least tendency to overfitting data. Another experiment was to build Decision Tree Classifier models with 'Gini' index criterion and tree depths ranging from 1 to 10. With Gini index, it is observed

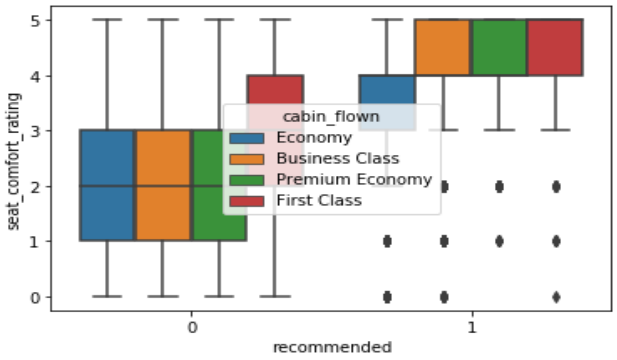

Fig. 6. Seat Comfort rating Vs Recommended

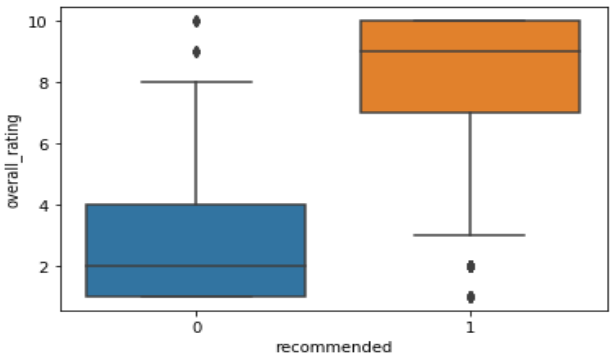

Fig. 7. Overall rating Vs Recommended

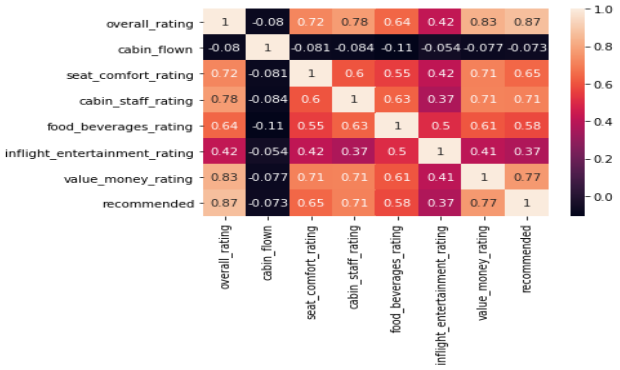

Fig. 8. Heatmap

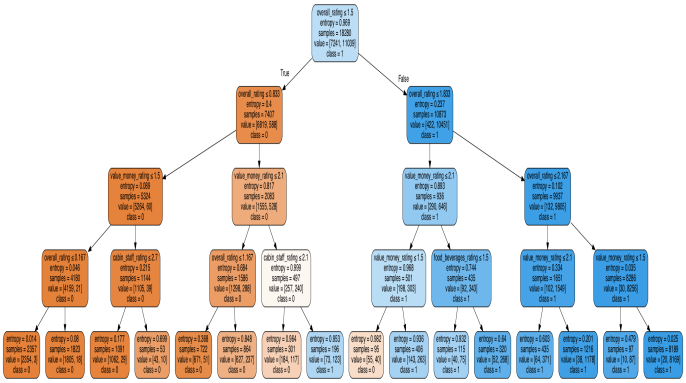

Fig. 9. Entropy Decision Tree with Tree depth = 4

that model with tree depth 4 and 5 gave the best accuracy scores. A decision tree image was built with the Entropy model of tree depth 4 using sklearn tree's export graphviz function to better visualize the Decision Tree model. Experiment was further continued with applying 10- fold cross-validation for decision tree classifiers to better understand the extent to which accuracy scores varied.

B. Results And Discussion

Fig. 9 shows Decision Tree plot on the Entropy model with tree depth 4. It can be observed from the decision tree that overall rating feature of the dataset is the root node, which implies it has the highest information gain. The next most important feature is value for money followed by cabin staff rating and food beverages rating on tree depth levels 2 and 3 of the decision tree. Class labels with values 0 and 1 are the leaf nodes of the decision tree which classify the dataset based on recommended target feature.

Accuracy of the above decision tree entropy model is 93.52 percent. Metrics derived from sklearn library including classification report provides details of accuracy, recall, precision and f1-score of the prediction model.

Fig. 10 shows that accuracy of the predictions of decision tree model is approximately 94 percent. Precision for classification ranges from 91 to 95 percent. Recall values of the decision tree prediction model ranges from 93 to 94 percent. f1-score ranges from 92 to 95 percent. Further experiment results of Decision Tree classifier with different criterion like 'Entropy' and 'Gini' with varying tree depth levels have accuracy scores as shown in Fig. 11 and 12.

Fig. 11 shows accuracy of Entropy decision tree models with tree depths 1 to 10 as 94 percent and decreasing to 93 percent respectively. Fig. 12 shows accuracy of Gini index decision tree classifier models ranging from 94.05 percent to 93.52 percent. In both the criterion models, it can be observed that models with tree depths 4 and 5 have highest accuracy of approximately 94.5 percent.

Fig. 13 and Fig. 14 shows the graph plots of accuracy scores of Entropy Decision Tree models and Gini index Decision Tree classifier models with tree depths from 1 to 10. The blue line represents score of Training dataset and Orange line represents score of Testing dataset. It can be observed that as the tree depth level is less, the model tends to overfit. As the tree depth level increments by 1, the model accuracy stabilizes at tree depth 4 and 5. While tree depth increases to 7 through 10, we can observe that the decision tree classifier model is under-fitting and accuracy range varies from 94.0 percent to 96.0 percent.

Fig. 15 shows the K-Fold Cross Validation accuracy scores of the Decision Tree Classification model with 'Entropy' criteria and tree depth 4. K-fold levels were applied from 1 to 10 and scores range from 85 percent to 92 percent.

Classification Report of Decision Tree Classifier Prediction model::

	precision	recall	f1-score	support
0.0	0.91	0.93	0.92	3567
3.0	0.95	0.94	0.95	5437
accuracy			0.94	9004
macro avg	0.93	0.93	0.93	9004
weighted avg	0.94	0.94	0.94	9004

Fig. 10. Classification Report of Decision Tree Classification model

Decision Tree 'Entropy' :::

	LevelLimit	Score for Training	Score for Testing
1	1.0	0.944748	0.940582
2	2.0	0.944748	0.940582
3	3.0	0.944748	0.940582
4	4.0	0.948304	0.946135
5	5.0	0.948687	0.945802
6	6.0	0.950821	0.943914
7	7.0	0.952735	0.945024
8	8.0	0.954923	0.943581
9	9.0	0.958807	0.942359
10	10.0	0.962582	0.938028

Fig. 11. Entropy Decision Tree Depth and Accuracy Score

Decision Tree 'Gini' :::

	LevelLimit	Score for Training	Score for Testing
1	1.0	0.944748	0.940582
2	2.0	0.944748	0.940582
3	3.0	0.945241	0.943247
4	4.0	0.947155	0.944025
5	5.0	0.949453	0.945247
6	6.0	0.951313	0.943692
7	7.0	0.954267	0.941359
8	8.0	0.957385	0.940027
9	9.0	0.960011	0.939138
10	10.0	0.964551	0.935251

Fig. 12. Gini Decision Tree Depth and Accuracy Score

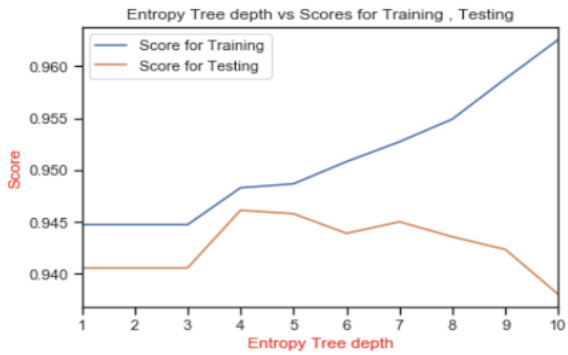

Fig. 13. Entropy Decision Tree Depth and Accuracy Score Graph

V. Random Forest Classification

As random forest consists of many decision trees created by selecting some random features k from given m features of dataset, and the one with the highest vote on test prediction is taken as a prediction result. It can be slow to make a prediction, as each decision tree has to perform a prediction, but to overcome this challenge we can limit the number of trees created, so for this dataset we use the number of trees to be 10.

A. Results And Discussion

By using the random forest classifier on the airline dataset, we are able to achieve an accuracy of 0.9365. The random forest classifier from sklearn library is used with the parameter of n-estimators (number of trees to be created) as 10 and random-stat 42, along with bootstrap being true. The train and test splitting of the dataset is performed using cross validation method.

VI. Naive Bayes

This classification technique based on Bayes Theorem, with the assumption that features are independent. Bayes theorem makes it possible to calculate the probability of a data point belonging to a class. It works by converting the data to frequency tables and than find the likelihood, followed by finding the posterior probability for each class. As the biggest challenge of this classification model is that it suppose the features to be independent, but in real life dataset it's not possible for the features to be independent, likewise our dataset consists of correlated features which is why the performance of this model is not as good as other classifier.

A. Results And Discussion

The naive bayes model from sciki-learn library is used for the dataset of airline. It has three options of Gaussian, Multinomial and Bernoulli. We are using Multinomial one for our dataset's classification. Tha accuracy of this model for our airline dataset is 0.822.

VII. Support Vector Machines

SVM classifier tries to classify the dataset's different classes by drawing a differentiating hyperplane between them. The hyperplane is selecting in a way that it could do the correct classification and it has the maximum margin from the data points of each class. For the dataset which is not linearly separable, SVM transform it to a higher dimensional space and than do the classification on this linearly separable transformed dataset. As the challenge of SVM is that it under performs for noisy dataset, so to overcome this issue, we did a data preprocessing of the airline dataset and removed the noise from it.

Fig. 14. Gini Decision Tree Depth and Accuracy Score Graph

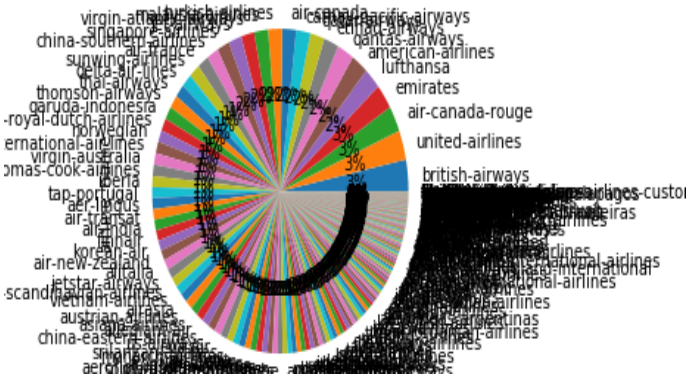

Fig. 16. Random Forest Model

```
K-Fold level =  1  Score =  0.8933675338951997
K-Fold level =  2  Score =  0.9274459508977647
K-Fold level =  3  Score =  0.9164529131550019
K-Fold level =  4  Score =  0.9245144741663613
K-Fold level =  5  Score =  0.9329424697691462
K-Fold level =  6  Score =  0.9248809087577867
K-Fold level =  7  Score =  0.9244868035190615
K-Fold level =  8  Score =  0.8834310850439883
K-Fold level =  9  Score =  0.9288595526219289
K-Fold level =  10  Score =  0.856985698569857
```

Fig. 15. 10-Fold Cross Validation Accuracy Scores of Decision Tree model

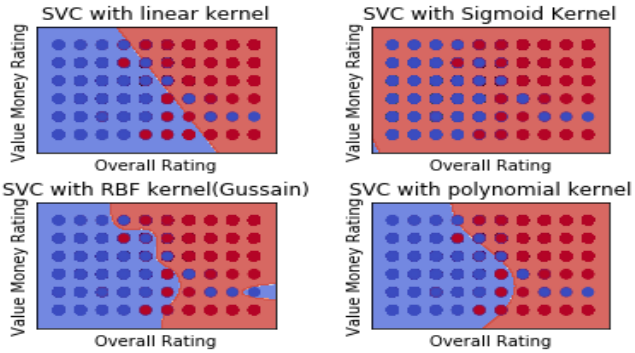

Fig. 17. SVM with different kernels

A. Results And Discussion

The SVM classifier from sklearn library is used for the airline dataset. As this method works with many available kernels (Gaussian, Sigmoid, Linear, Polynomial), so we tried to run the model with each kernel on our dataset and evaluated the accuracies. The accuracy of Linear, Polynomial and Gaussian kernels are observed to be the same, which is 0.94, while the Sigmoid kernel's accuracy is only 0.31. Each kernel tries to draw the classification hyperplane differently, as shown below,

The other parameters of SVM are gamma and C. Gamma refers to the coefficient of rgb (Gaussian), poly and sigmoid kernels, while C controls the trade off between smooth decision boundary and classifying the training points correctly. Increasing both this paramters would cause overfitting of the model, as shown in figures below, so for our dataset we chose gamma = 0.7 and C=1.

VIII. Clustering

Clustering is grouping set of objects based on characteristics and aggregating them according to their similarities. It is a Machine Learning technique that involves the grouping of data points. The whole goal is to extract information from the dataset and transforming it into understandable form.

A. PCA

Principal Component Analysis (PCA) is an unsupervised machine learning technique.It is a popular technique for deriving a set of low dimensional features from large set of variables.

B. K-Means

The process on how k-means works is first we select a number of classes/groups to use. Then we randomly initialize their respective center points.Later each data point is classified by computing the distance between that point and each group center, and then classifying the point to be in the group whose center is closest to it. Continued by Recomputing the group center by taking the mean of all the vectors in the group.Finally repeat these steps for a set number of iterations or until the group centers don't change much between iterations.

Advantages: - K-Means is pretty fast. - linear complexity O(n*n)

Disadvantages: - Select number of groups/classes in advance. - Results may not be repeatable and lack consistency as we randomly select centroids.

C. Results

From Fig 19 we can see that there is a clear distinction between the two clusters. The Blue points represent recom-

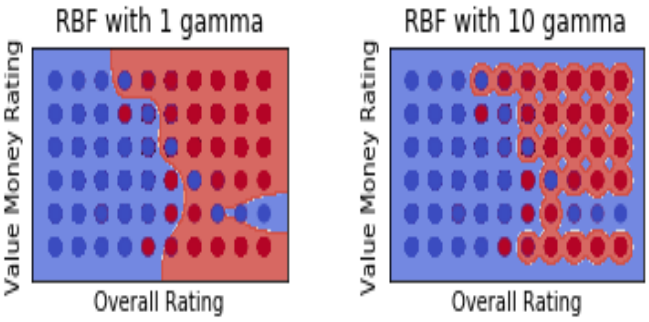

Fig. 18. SVM with different Gamma values

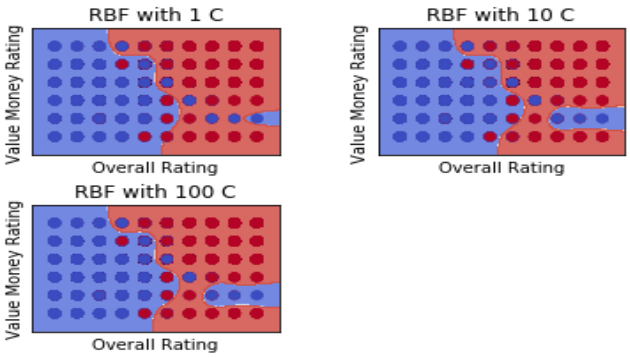

Fig. 19. SVM with different C values

mended where as the red represents not recommended. Even though the recommended points are dominant we can see the two clusters clearly.

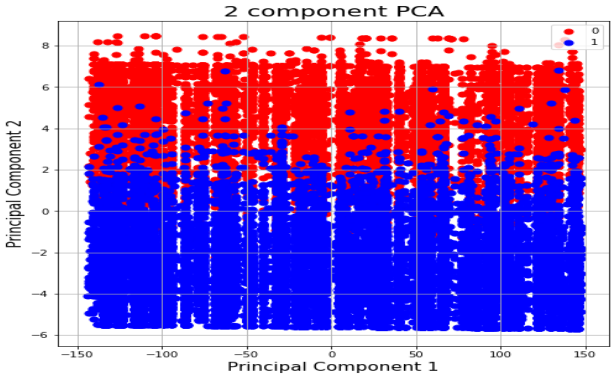

Fig. 20. PCA Principal Components

If we observe the inertia plot in Fig 20 there's a sharp dip when the number of clusters is 2. Which re-confirms the fact that the algorithm for this particular dataset works best when K=2.

IX. CONCLUSION AND FUTURE WORK

In this study, we were able to successfully apply machine learning algorithms to predict whether an airline is recommended or not based on the airline ratings and show classifiers like Decision Tree Classifier, Random Forest Classifier and Support Vector Machines can predict if an airline is recommended or not fairly accurately. For future work we like to further improve our models, perhaps with more training-data or deep neural network, or both.

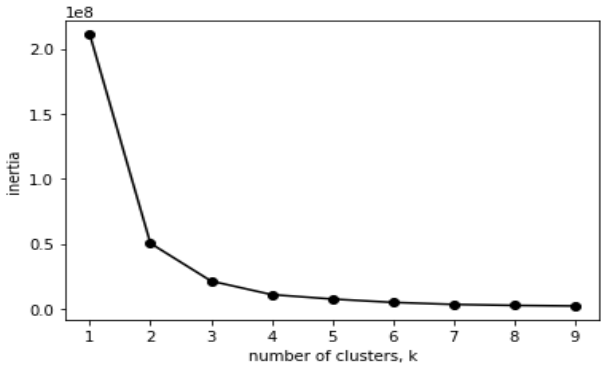

Fig. 21. Inertia Vs No of Clusters

REFERENCES

[1] "Airline dataset mining," accessed: 2019-11-20.