University of
Strathclyde
Glasgow

CS982: Big Data Technologies

Report on Airline Passenger Satisfaction survey

Contents

Chapter 1 :Introduction & Problems addressed: 3

Chapter 2 :Dataset- Airline Passenger Satisfaction 4

Chapter 3 :Exploratory Data Analysis and Summary 6

Chapter 4 :Unsupervised Analysis 12

Chapter 5 :Supervised Analysis 14

Chapter 6 :Reflections & Conclusion 16

Appendix 18

Bibliography 19

Figures:

Figure 1-1 Analysis by McKinsey & Company 3

Figure 3-1 Proportion of Satisfied and Unsatisfied People 6

Figure 3-2 First-Time vs Returning Passenger Satisfaction Levels 7

Figure 3-3 Satisfaction levels of Different Passenger Class passengers 8

Figure 3-4 Satisfaction levels of different Genders 9

Figure 3-5 Age Distribution of passengers 9

Figure 3-6 Best and Worst services 10

Figure 3-7 Flight Distance vs Arrival Delay 11

Figure 4-1 Inertia Vs Number of clusters 13

Figure 5-1 Confusion matrix 16

Tables:

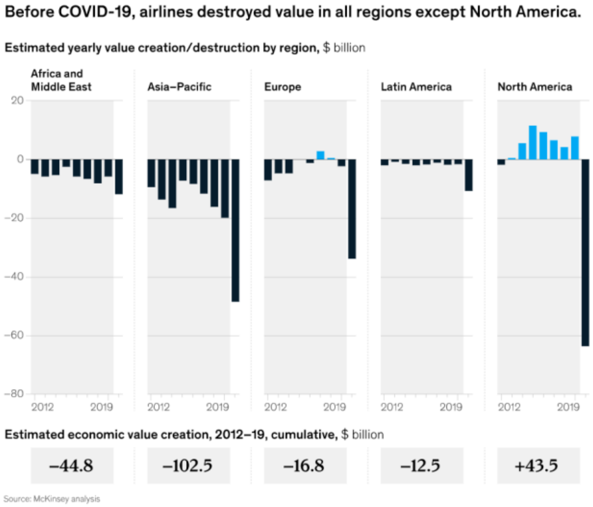
Table 5-1 Classification report 15

Table 5-2 Confusion matrix 15

Chapter 1 :Introduction & Problems addressed:

The airline industry has seen tremendous growth over the past few decades, and before the pandemic, the industry served 4.5 billion passengers on scheduled services (*The World of Air Transport in 2019*, 2020). This upward trend saw major changes due to the recent Covid19 outbreak, during which the airline industry suffered massive losses. As per some estimates, the industry suffered a loss of $372 billion in 2020 alone (*Effects of Novel Coronavirus (COVID-19) on Civil Aviation: Economic Impact Analysis Economic Development-Air Transport Bureau*, 2022).

However, it is interesting to note that even before the pandemic a number of Airline companies including Kingfisher airlines, Transaero airlines and many more had either filed for bankruptcy or shut down their operations after incurring huge losses in their business (*Famous Airlines That Have Gone Out of Business*, 2020). The airline industry is constantly in news for its low capital returns and is often considered a non-profitable sector among investors. Figure 1-1 below gives a better illustration of how unprofitable the airline industry has been(*COVID-19's impact on the global aviation sector | McKinsey*, 2022)


Before COVID-19, airlines destroyed value in all regions except North America.

Estimated yearly value creation/destruction by region, $ billion

Source: McKinsey analysis

Figure 1-1 Analysis by McKinsey & Company

In order for an airline company to survive in such a market, they have find ways to outcompete it's business rivals. They could focus on improving their customer experience by conducting regular surveys of their passengers and understanding where there is room for improvement. Numerous airline companies may have already adapted to using Big data technologies and data analytics to find interesting trends among passengers, and also find useful patterns contained within the data to segregate different types of passengers. The results of such a study can then be used to improve their services and also create personalized services for different categories of passengers.

This report is a detailed analysis of the passenger satisfaction survey results of an anonymous airline company. To begin with, using exploratory data analysis we will try to understand the demography of passengers and the relation between various parameters. We will also identify which service offered by the aircraft is disliked by most passengers, and which service was highly rated by most passengers. Next, we will use KMeans unsupervised machine learning algorithm to find interesting clusters that are formed using ratings corresponding to different services in the dataset. Finally, we will develop a supervised model to predict customer satisfaction based on a number of parameters present in the dataset.

Chapter 2 :Dataset- Airline Passenger Satisfaction

The dataset used for this report is publicly available and can be downloaded from the Kaggle website using the following link: https://www.kaggle.com/datasets/mysarahmadbhat/airline-passenger-satisfaction.

This data is likely to be from an anonymous airline company. The dataset itself contains the results of a survey that had 129880 participants. As a result, the dataset contains 129880 rows and 24 columns. The first nine columns describe the nature of passengers and the flight they traveled in, they are:

- ID: Unique ID for each passenger
- Gender: Gender of each passenger, (Male/Female)
- Age: The age of each passenger
- Customer Type: Indicates whether it is a first-time or Returning Customer
- Type of Travel: Purpose of Travel (Personal/Business)
- Class: Travel class in the airplane for the passenger seat.

- Flight Distance: Flight distance in miles
- Departure Delay: Delay in the departure expressed in minutes
- Arrival Delay: Delay in the arrival expressed in minutes

The subsequent 14 columns indicate the passenger satisfaction ratings received for 14 different services offered by the airlines, on a scale of one to five (zero means, the service is not applicable to a specific passenger type). The 14 services are as follows:

- Departure and Arrival Time Convenience
- Ease of Online Booking
- Check-in Service
- Online Boarding
- Gate Location
- On-board Service
- Seat Comfort
- Leg Room Service
- Cleanliness
- Food and Drink
- In-flight Service
- In-flight Wifi Service
- In-flight Entertainment
- Baggage Handling

The final column “**Satisfaction”** indicates the overall satisfaction of the traveling experience for each passenger with the airline. It has two values, the first one is “Satisfied” and the second one is “Neutral or Dissatisfied”.

As part of the preprocessing step before exploratory data analysis, the data was checked for the presence of null values first. The ‘Arrival Delay’ column was found to have 393 null rows and they were filled with zero. Next, the dataset for checked for the validity of non-null values. The columns Gender, Customer Type, Type of Travel, Class, and Satisfaction were found to have distinct values as expected and no junk values were present. For the age column, there were no junk numeric values lesser than 0 and greater than 100. This concludes the process of data cleaning.

Chapter 3 :Exploratory Data Analysis and Summary

Out of the 129880 participants that took part in the survey only 43% of the passengers were satisfied with their travel experience with the airlines, the rest were either neutral or dissatisfied. It is illustrated in the Figure 3-1 below:

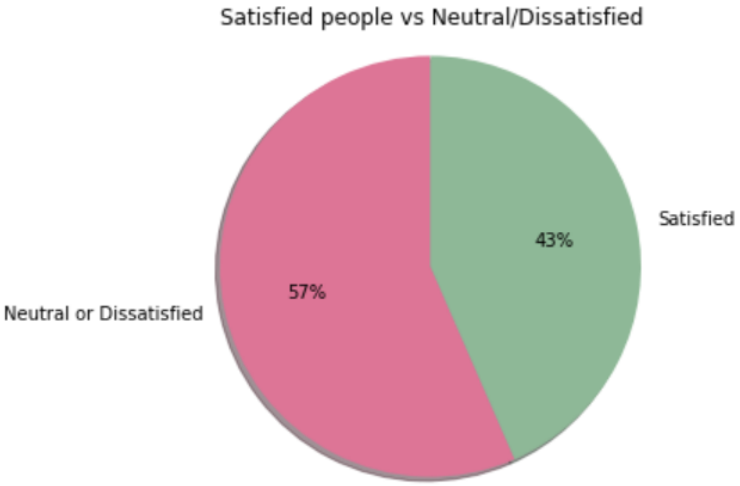


Figure 3-1 Proportion of Satisfied and Unsatisfied People

When less than half of the passengers are satisfied with the airlines, the customer retention rate of the airlines is bound to be low, because the dissatisfied customer may choose another airline operator for their future trips. We can observe the relationship between the Customer Type and their corresponding Satisfaction in figure 3-2 below:

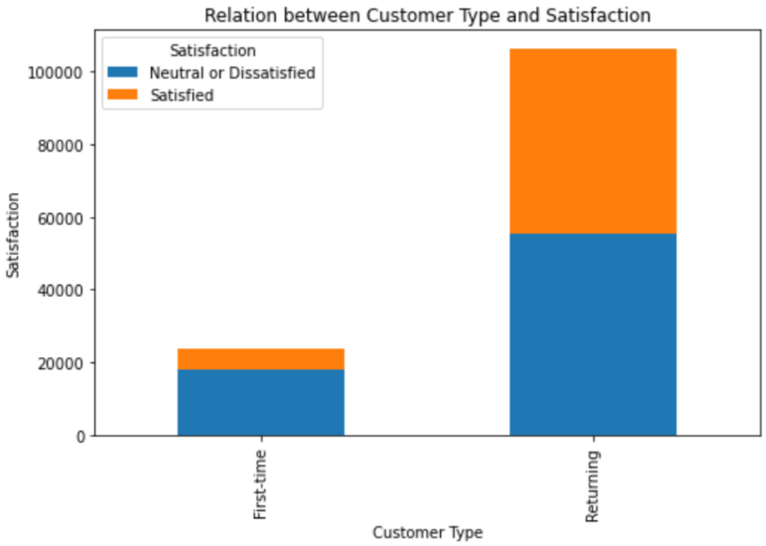


Figure 3-2 First-Time vs Returning Passenger Satisfaction Levels

We can notice that among the returning passengers, the number of satisfied passengers in comparison to neutral or dissatisfied passengers is more or less the same. However, among first-time travelers the level of satisfaction is low. This means, that the airline must improve the travel experience of first-time travelers with the aircraft. We can also infer that the returning passengers, like certain services provided by the airlines, which is why they are reusing the services of these airlines. The airlines must ensure that the quality of such services does not degrade over time

Next, we shall examine the relationship between passenger class and satisfaction. We can observe from figure 3-3 that business class passengers are among the most satisfied passengers in comparison to fellow passengers from other classes.

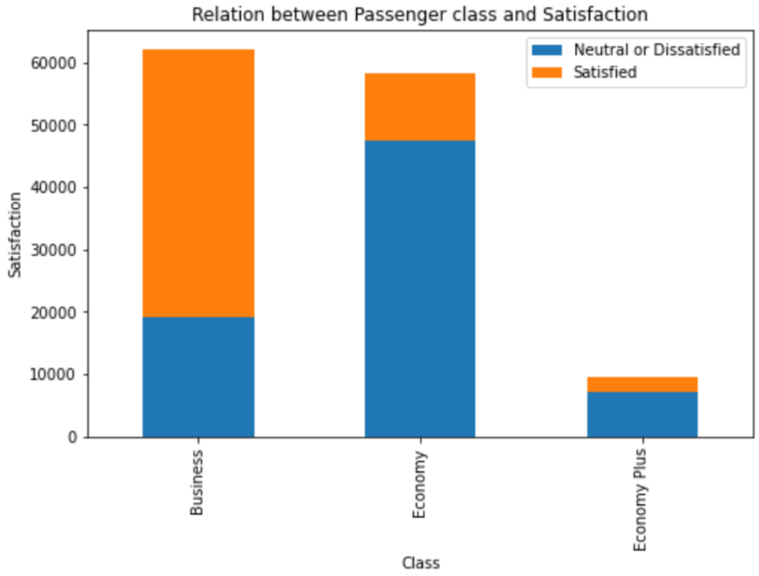


Figure 3-3 Satisfaction levels of Different Passenger Class passengers

This could mean that certain services offered exclusively to business class passengers are keeping them happy, while such services may neither be applicable nor be of the same quality to passengers of other classes. In such cases, the passengers of economy classes may have had a neutral experience with respect to certain services.

Besides this, we will also examine the level of satisfaction among passengers of different genders to understand if the quality of services is biased towards any gender. We can notice from Figure 3-4 that the satisfaction levels of both sexes are almost identical which would mean that all the services are unbiased towards a specific gender.

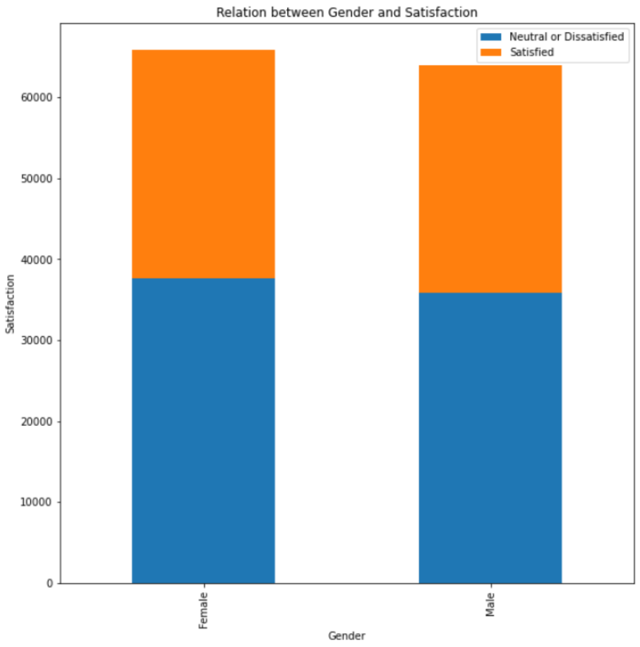


Figure 3-4 Satisfaction levels of different Genders

Moving on, we will try to understand the demography of our passengers, what is the minimum age, maximum age, and the average age of passengers flying on this airline. From figure 3-5 below we can see that the minimum age was 7 years and the maximum age was 85 years and the mean age was 39.4.

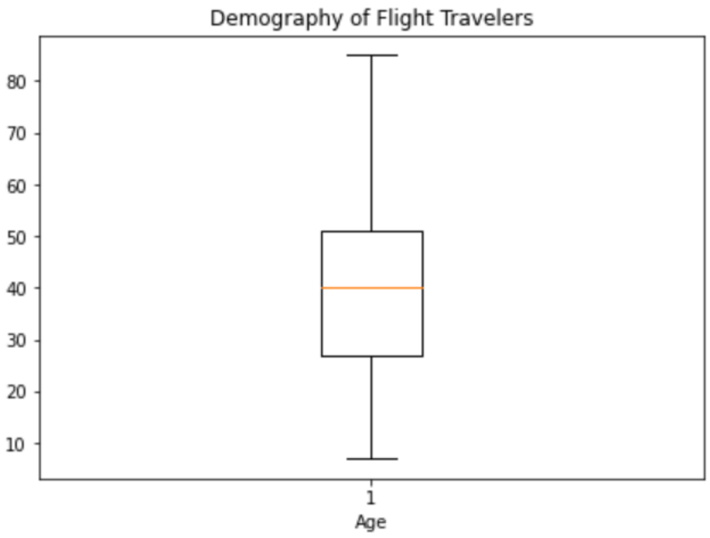


Figure 3-5 Age Distribution of passengers

Such information will give us on idea whether additional on-board services for infants and mobility services for elderly passengers are required on the flight.

Continuing our analysis, we will next try to understand the best and the worst of airlines' services. One way to know the best and worst service is by examining which services received the highest 5-star and 1-star ratings respectively. This is better illustrated by Figure 3-6 below:

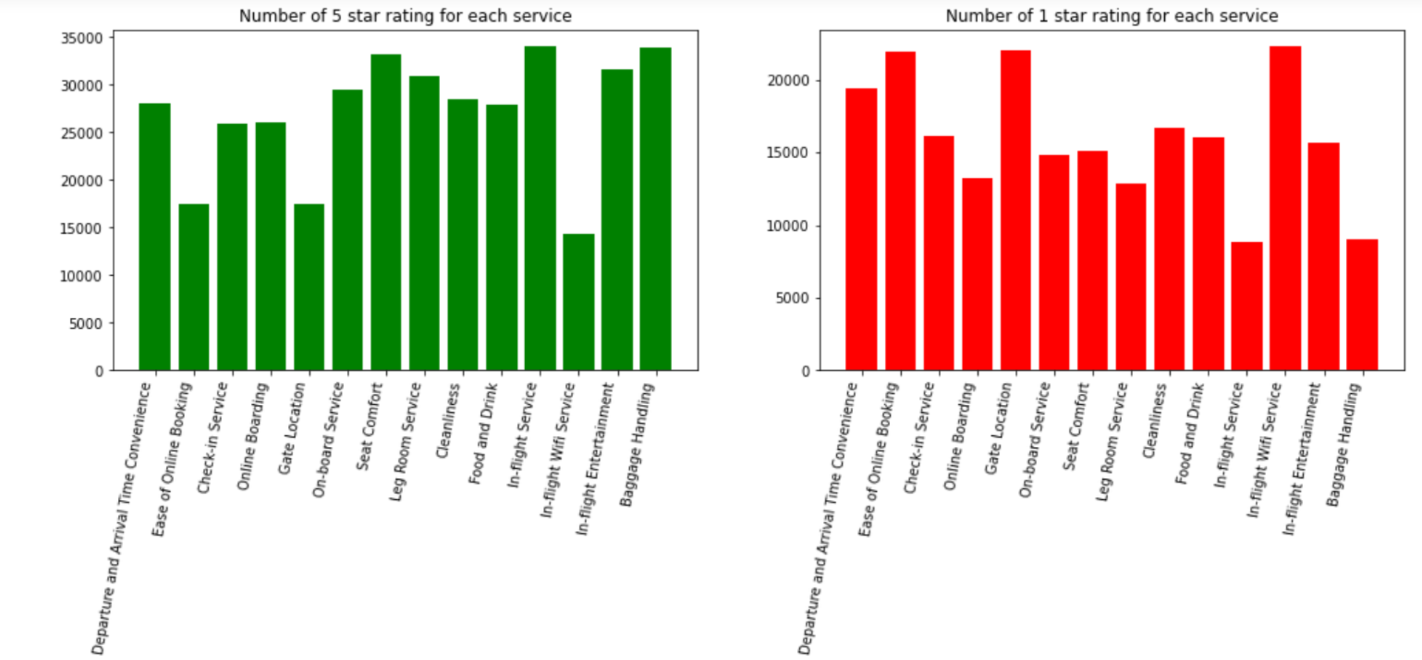


Figure 3-6 Best and Worst services

From the above figures we can clearly identify that the top 3 best services are In-flight service, followed by baggage handling and Seat Comfort. The airlines have to ensure that the quality of these services remains the same over time. From the same figure, we can also identify the top 3 worst services as In-Flight Wi-Fi Service followed by Gate location and Ease of Online Booking. The airline clearly needs to address the quality of In-flight Wi-Fi service by switching to a better Internet Service Provider. They also need to spend time on improving the User Experience on their mobile application and website to ensure that booking tickets online is hassle-free.

Another important observation from Figure 3-6 is that the Departure and Arrival Time convenience has been rated low by many passengers as well. There could be numerous factors such as weather conditions, availability of a strip on the runway for takeoff and landing, loading of baggage onto the flight, etc. that could lead to departure and arrival delays , is flight distance one of those factors? We shall examine this by trying to find a relation between flight distance and arrival delay as shown in Fig 3-7.

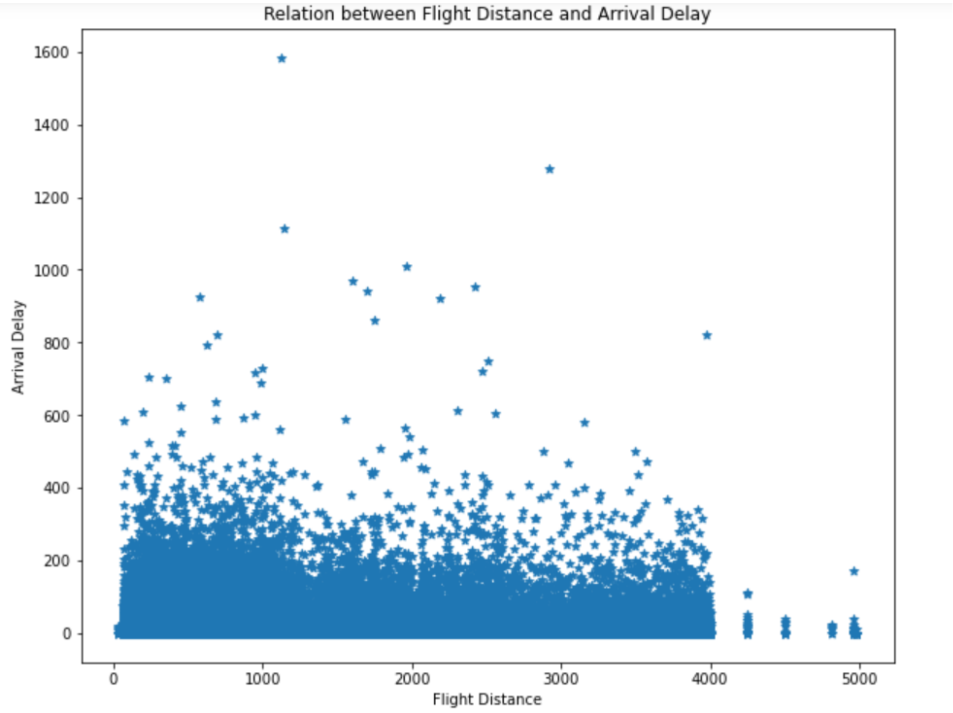


Figure 3-7 Flight Distance vs Arrival Delay

From the figure we can infer that there is no direct correlation between the flight distance and arrival delay.

Chapter 4 :Unsupervised Analysis

According to Bandyopadhyay, Sanghamitra, Classification is the task of assigning labels to a set of instances in such a way that instances with the same label share some common properties and logically belong to the same class. (Bandyopadhyay and Saha, 2013, p.1) . When preclassified, labeled instances are not available, the task of classification is often referred to as unsupervised classification or clustering. (Bandyopadhyay and Saha, 2013, p.1). Clustering is a great tool for data analysis, customer segmentation, recommender systems, search engines, image segmentation and many more(Geron and Aurelien, 2019).

We will use KMeans clustering to identify clusters formed among the various columns related to the rating of different services offered by the airlines. Although we already know there are two Satisfaction classes existing in the dataset, it would be interesting to find out if there are other meaningful groups formed based on the different rating attributes. For KMeans algorithm to work we have to supply the value for 'K' to create K clusters. The features which we will be using for this model are all the columns related to the rating of different services, scaled on a standard scaler. In order to identify the optimal value for K, we plot the model's inertia, which is the mean squared distance between each instance and it's closest centroid, as a function of the number of clusters, as suggested by Geron and Aurelien (Geron and Aurelien, 2019,p.246). We end up with a graph shown in fig 4-1 below:

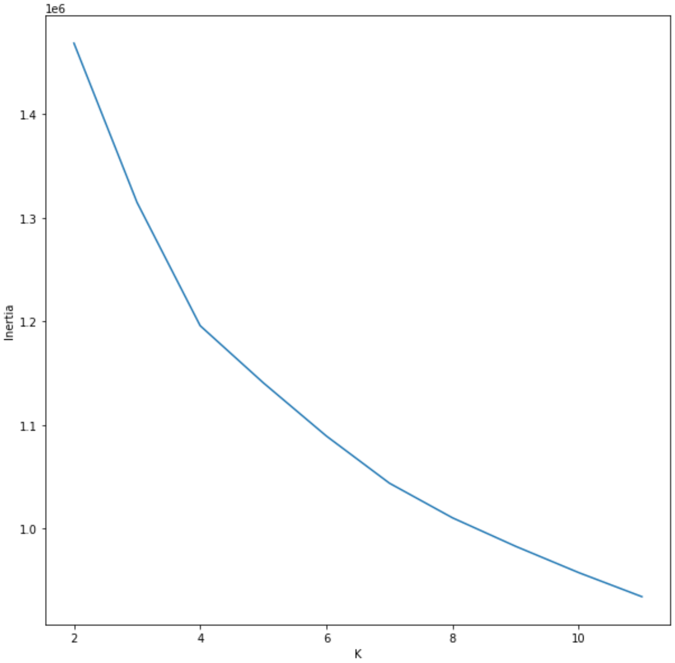


Figure 4-1 Inertia Vs Number of clusters

We can notice an inflexion point(similar to the elbow of our hand), corresponding to the value of K=4, this would be a good value for K as per the elbow method.

We shall next compare the silhouette score for clusters starting from 2 clusters until 6 clusters, because we know from figure 4-1 that an optimal value lies between 2 and 6. As defined by Geron and Aurelien, “The silhouette score, is the mean silhouette coefficient over all instances. An instance’s silhouette coefficient is equal to (b-a)/max(a,b), where a is the mean distance to the other instances in the same cluster and b is the mean nearest cluster distance” (Geron and Aurelien, 2019,p.24). For our model, we can notice that the silhouette score for 2 clusters is highest with a value of 0.17. Clusters 3 and 4 also have a considerably high values of 0.15 and 0.14 respectively. Finally, we notice that the model with 2 clusters has the highest completeness score of 0.17. A clustering result satisfies completeness if all the data points that are members of a given class are elements of the same cluster(*sklearn.metrics.completeness_score — scikit-learn 1.1.3 documentation,* no date). We can hence conclude that the optimum number of clusters is 2, this corresponds to the actual number of classes that exist for the Satisfaction column in the dataset.

Chapter 5 :Supervised Analysis

Supervised learning algorithms try to *model relationships and dependencies between the target prediction output and the input features* such that we can predict the output values for new data based on those relationships which it learned from the previous data sets(Types of Machine Learning Algorithms You Should Know | by David Fumo | Towards Data Science, no date).

I will build a model to predict the satisfaction level of each passenger based on how they rated the different airline services. The independent variables used for this model are all the columns related to ratings for different services and the dependent variable is the satisfaction column. The dependent variable is a string and we will convert it into numeric values by mapping "Satisfied" to integer 1 and "Neutral or Dissatisfied" to integer 0. Next, I have split the training and test data in a proportion of 80:20 using the model_selection.train_test_split function. Next, I move on to training the model using a suitable algorithm. After carefully examining the accuracy score of Logistic Regression, Decision Tree, K Nearest Neighbors, Naive Bayes Gaussian, and Naive Bayes Multinomial prediction models, for the purpose of this report, I will be using Decision Tree model which was found to have the highest accuracy score. The accuracy score indicates how well the set of labels predicted matches the actual labels of the target test data(Y_test).

I have next built the prediction model using the Decision tree classifier. According to Carolina Bento, "The intuition behind Decision Trees is that you use the dataset features to create yes/no questions and continually split the dataset until you isolate all data points belonging to each class. With this process, you're organizing the data in a tree structure."(Carolina Bento, 2021)

Once the model is trained, I then predict the labels using the test data. To determine the accuracy of the model I use metrics such as score, accuracy score, classification report, and confusion matrix. The classification is presented in table 5-1 below

	precision	recall	f1-score	support
	0.94	0.94	0.94	14730
	0.91	0.91	0.91	11246
accuracy			0.93	25976
macro avg	0.93	0.93	0.93	25976
weighted avg	0.93	0.93	0.93	25976

Table 5-1 Classification report

- The precision indicates, out of all “satisfied” passenger predictions, how many were actually right, which is (correct prediction of the satisfied passengers)/(Total satisfied class prediction) i.e, True Positive / (True Positive + False Positive)
- Recall indicates, out of the total Satisfied passengers in the sample, how many were predicted correctly, which is (correct prediction of satisfied passenger)/ (total satisfied passengers in actual sample) i.e (True Positive/(True Positive + False Negative).

The analogy of True Positive, True Negative, False Positive and False Negative can be visualized in a confusion matrix that can be determined by using the metrics.confusion_matrix function. It is shown in table 5-2 below:

False Positive	False Negative
[[13622	987
935	10432]]
True Negative	True Positive

Table 5-2 Confusion matrix

For better visualization it can be plotted on a heat map as shown below in Figure 5-7

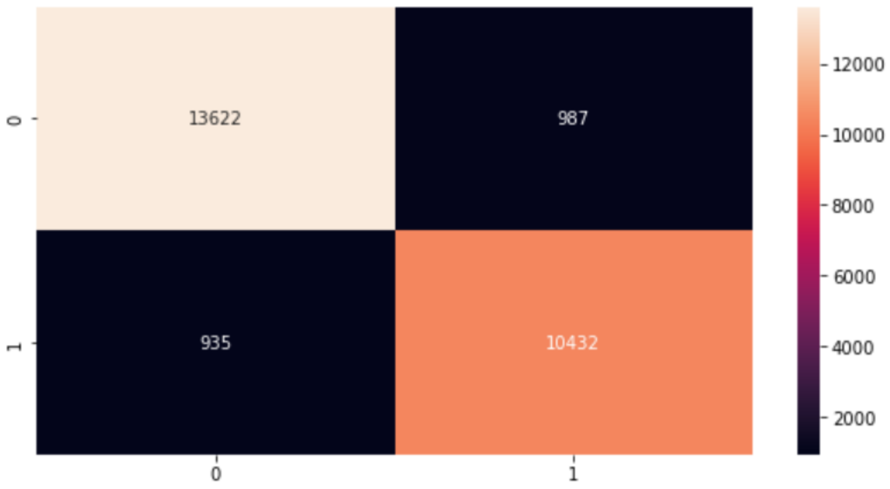


Figure 5-1 Confusion matrix

Chapter 6 :Reflections & Conclusion

The dataset was well suited for exploratory data analysis and building prediction models using supervised analysis. However, it was not good for unsupervised learning. When trying to find relation between age, passenger class, gender with one of the airline services, the resulting scatter plot was had distinct straight lines with data points on a single line, plotted against the distinct categorical value or rating values. Hence, such features were not ideal for clustering to identify some underlying relationship between the columns. This meant that I had use all the rating columns to identify clusters and hidden trends. However, as there were 14 different columns, the clustering process took time, and the computation of the silhouette score took longer.

Nevertheless, the prediction model developed using the Decision Tree analyzer was fairly accurate. Some degree of inaccuracy is because the satisfaction class is imbalanced. i.e, the number of satisfied passengers is lesser than that of neutral or dissatisfied passengers. Although during my analysis, the K Nearest Neighbours model was comparatively more accurate than the Decision tree model, the former was slower and took long time to give an output. This is because as the dataset is large, I had to use a large value for k (value that determines how many nearest neighbours to be considered).

Finally, I was able to find interesting trivia about the dataset using techniques taught in class and the results of exploratory data analysis was fairly satisfactory.

Appendix

Software versions & Packages used:

Python version – 3.9.12

IDE: Jupyter Notebook : 6.4.8

Packages used:

- Pandas
- matplotlib.pyplot
- seaborn
- metrics from sklearn
- cluster from sklearn
- LabelEncoder from sklearn
- Scale from sklearn
- Selection from sklearn
- DecisionTreeClassifier from sklearn.tree

Reference Manager: Mendeley Reference Manager

Bibliography

Bandyopadhyay, S. and Saha, S. (2013) *Unsupervised classification: Similarity measures, classical and metaheuristic approaches, and applications*, *Unsupervised Classification: Similarity Measures, Classical and Metaheuristic Approaches, and Applications*. Springer-Verlag Berlin Heidelberg. doi:10.1007/978-3-642-32451-2.

Carolina Bento (2021) *Decision Tree Classifier explained in real-life: picking a vacation destination | by Carolina Bento | Towards Data Science*. Available at: https://towardsdatascience.com/decision-tree-classifier-explained-in-real-life-picking-a-vacation-destination-6226b2b60575 (Accessed: 2 November 2022).

COVID-19's impact on the global aviation sector | McKinsey (2022). Available at: https://www.mckinsey.com/industries/travel-logistics-and-infrastructure/our-insights/taking-stock-of-the-pandemics-impact-on-global-aviation (Accessed: 1 November 2022).

Effects of Novel Coronavirus (COVID-19) on Civil Aviation: Economic Impact Analysis Economic Development-Air Transport Bureau (2022).

Famous Airlines That Have Gone Out of Business (2020). Available at: https://www.businessinsider.com/airlines-that-go-out-of-business-2019-3?r=US&IR=T#transaero-defunct-2015-17 (Accessed: 1 November 2022).

Geron and Aurelien (2019) 'Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems', in. O'REILLY.

sklearn.metrics.completeness_score — scikit-learn 1.1.3 documentation (no date). Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.completeness_score.html (Accessed: 2 November 2022).

The World of Air Transport in 2019 (2020). Available at: https://www.icao.int/annual-report-2019/Pages/the-world-of-air-transport-in-2019.aspx (Accessed: 1 November 2022).

Types of Machine Learning Algorithms You Should Know | by David Fumo | Towards Data Science (no date). Available at: https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861 (Accessed: 2 November 2022).