Airline Customer Satisfaction

Group Members:

Y Nikhil Bhardwaj	P Sai Varshith	P J Subramanya Hande
PES1UG19CS586	PES1UG19CS320	PES1UG19CS316

1) Introduction and background – what is the problem area? Why is it important? What is the specific problem you seek to solve

Competition is the first word which will come to one's mind when they want to do any business or any other domain of interest.

There are many modes of travel like railways, roadways, waterways and airways where the airways are the costliest even if there is a great demand for it. With demand competition also increases.

Airline administration and management comes under the service domain where customer satisfaction and trust are the driving key for success of the organisation. Many airline companies collect reviews from customers for the scope of improvement.

This feedback helps us to get insights in which area they can improve to get more profits. The data given is about airline organization which contains the details of the customers used by their airlines and their rating in different aspects. In simple words, customer satisfaction is a measurement that determines how well a company's products or services meet customer expectations and requirements.

It is important in understanding the user feedback in efficient manner for the better understanding of the company success since there is competition in every domain if we miss a chance, it is an opportunity for others to capture so we need to very precise with our advantages and disadvantages of services

In our data set, we are collecting information from customers who travelled through their airlines and collected information regarding timing, delay, service onboard entertainment, service quality, Food and drink, hospitality and so on. Positive feedback is the expectation of any public domain organisation.

We collected feedback from customers in the form rating which ranges from 0-5 where 0 is considered as bad ,1 is needed to improve,2 is average,3 is ok,4 is good and 5 is excellent.

Customer satisfaction scores indicate how the firm is going to perform in future.

Identifying and addressing the problems of customers, especially dissatisfied customers is considered as crucial because it shows that we are paying our attention towards their interest and respecting their opinion. This makes more chances for customers to come back and carry on with us for a long period which is our ultimate goal as a service firm. There are many customers who do not give the review properly and data may be missing many times (i.e.; incomplete data) which need to be scaled to get the correct insights.

We will get the insights where to improve if we improve those parts of the organisation their will be more success for the firm.

2) Previous work – A brief review of only the most relevant predecessor work; what limitations have you identified that you seek to address in your work? What are the assumptions you have made about the data/ problem area or the scope of the problem you seek to solve?

We are just keeping the references of different papers we found on our dataset so that we can better explain the limitations we found out in their work

a) LCA(Low Cost Airlines) providers always find ways to prove to their customers that it offers low price but not low quality. Many studies have been explored on the subjects of the service quality, cost, and customer satisfaction in the LCA around the world but still rarely any related research is in Vietnamese. Chen (2008) has appointed that customer satisfaction is a holistic concept that it represents the overall emotional response after consumption, and it can range from the level of dissatisfaction to satisfaction. LCA in particular, customer satisfaction is a crucial factor of competitive advantage and helps to create the success for LCA who is a new participant (Kim & Lee, 2011). Besides, customer satisfaction is a penchant for the defence to retain customers rather than for the offence to find new customers in a competitive environment (Reichheld & Sasser, 1990). Within theories about the elements of customer satisfaction.

b)It seems that price and service factors are core elements that influence strongly on customer satisfaction of a service company. In developing countries such as Vietnam, LCAs are always attractive to customers through cheaper fares due to low cost strategies and diversified revenue sources.

c)Many marketing studies (e.g., Overholt et al.,2007). revealed that there is certainly a positive link between high customer satisfaction from service quality and customers' loyalty. Indeed, there are many benefits from high standards of service quality such as sustainable demand from satisfied consumers as well as a positive image at the market. J. Chen and C. Gursoy (2001) stated that consumer loyalty in tourism can be traced back to high standard of quality According to S. Shaw (2007), more domestic passengers are satisfied with the quality of service, the more domestic passengers would become loyalty to a particular airline brand. Therefore, Airline industry should work on improving their standards of service quality in order to gain reputation and increase the overall level of satisfaction. They planned to use stratified sampling and random sampling techniques with a more diverse group of tourists and try to increase the variety of sample size provided time and budget allows

This study investigates the customer satisfaction of airline passengers and introduces perceived safety as a satisfaction driver, which has not yet been considered in the literature. Applying structural equation modelling to data collected from a sample of airline passengers reveals that perceived safety is one of the key drivers that can explain the degree of overall customer satisfaction

d)The limitations we found in these articles are that they were more focused on one factor such as low cost, services offered ,quality of services,safety, etc. But , we took more than one attribute into consideration and tried to classify the customer satisfaction based on all these attributes.
We removed some attributes which were not required in the dataset and won't matter much in the customer satisfaction through correlation analysis.

3) Proposed solution – an overview of the various components of your solution (preprocessing + building a model + evaluation)\
4) Experimental results and a detailed explanation of all the insights you have gained into the data (on what cases does the model work well? When does it fail?)

Preprocessing

We have encoded various attributes in our dataset such as satisfaction where there are 2 classes to mention whether the customer is satisfied or not based on other attributes values , And converted all of the data into numerical data from the categorical data wherever required so that analysis and prediction can become much easier , We have also done correlation analysis and found out that there are some redundant attributes that are not required which were 'Gate Location' and 'Type of travel' and hence we removed these both
We replaced NaN values in attributes like 'Arrival delay in minutes' with the median of the values of the column.
We have used various visualization techniques such as:
1) Bar graph to understand how 2 attributes vary w.r.t each other and we have even combined 2 attributes in some cases to see how they vary with satisfaction which gives a lot of information regarding how much these attributes matter in terms of customer satisfaction

2) Pie plot to analyse how all the attributes matter in terms of satisfaction i.e in pie plot we can see that customers are more interested in attributes like 'onboard services' and are less interested in attribute 'Food and drink'

Model Building

After preprocessing the data, we needed to build a model to classify if a customer is 'satisfied' or 'dissatisfied' based on all the different attributes and their values. This is a classification problem, hence we made use of some of the prominent classification models, starting with :

1) KNN model

KNN is one of the simplest ML algorithm based on supervised learning technique which assumes the similarity between user cases and put into most similar categories

We can assume k value based on the elbow method

In our model we found that the optimal k value is 5 after this the accuracy is reaching saturation

This model fails when k is less than 4 and it reaches to saturation after 7

```
X = df[['Gender', 'Customer Type', 'Age',
       'Class', 'Flight Distance', 'Seat comfort',
       'Departure/Arrival time convenient', 'Food and drink',
       'Inflight wifi service', 'Inflight entertainment', 'Online support',
       'Ease of Online booking', 'On-board service', 'Leg room service',
       'Baggage handling', 'Checkin service', 'Cleanliness', 'Online boarding',
       'Departure Delay in Minutes', 'Arrival Delay in Minutes']]
y = df['satisfaction']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state=42)
knn = KNeighborsClassifier(n_neighbors=3, metric='euclidean')
knn.fit(X_train, y_train)
y_pred = knn.predict(X_test)
confusion_matrix(y_test, y_pred)
metrics.accuracy_score(y_test, y_pred)
```

```
0.7014551894056051
```

The model accuracy of the KNN is very less. We know knn algorithm is very sensitive to outliers as we know some customers don't consider rating as important so they give some random rating so it is more prone to errors so the knn algorithm gave very less accuracy when compared to all other models

Accuracy of KNN Model for the optimal K value is 5

2) Logistic Regression:

The next model which we explored is Logistic regression which is also a supervised learning algorithm it provides probabilities for the understanding of the classification which makes it better in analysis of many datasets in our case it is binomial logistic regression

In this we splitted our dataset into training and test data set and made analysis by removing the class label for the analysis

The testing of the dataset is as

```
log_reg = LogisticRegression()
fit_model = log_reg.fit(X_train, y_train)
preds = fit_model.predict(X_test)
probs = fit_model.predict_proba(X_test)
```

```
true_count = 0
for pred, real in zip(model_results['PredictedClass'], model_results['TrueClass']):
    if pred == real:
        true_count = true_count + 1
print("Number of True Classifications = {0} ".format(true_count))
print("Accurate Classification Ratio = {0} ".format(true_count / len(y_test)))
```

```
Number of True Classifications = 31731
Accurate Classification Ratio = 0.7403233708966193
```

The number of True test cases when the test data percentage is 0.33 is 31731 which are correctly classified and predicted

The accuracy is less because in our dataset there are multiple decision boundaries

3) Naive Bayes Classifier

The Naïve Bayes algorithm is also a supervised algorithm where we split the dataset the same as in the Logistic regression algorithm It mainly comes to rescue when there is a high dimensional data.

In this particular we consider different features are independent i.e.; onboard entertainment is independent of duration of journey, Food quality is independent of class of ticket and so on which is the assumption of this algorithm it is one of the efficient algorithm because it reduces the dependencies and solve the problem in quick succession we used gaussian distribution for analysis for this algorithm

The accuracy obtained by this algorithm is 81.2% which is better than logistic regression and KNN

algorithm the accuracy is still less so explores through different other algorithms

The accuracy is low because it considers the features are independent but, in our model, there are some attributes which are related like flight distance and duration are related, Seat comfort Inflight service are related to Duration so the accuracy of this model is not high

```
X_train, X_test, y_train, y_test = train_test_split(X_data, y_data, test_size=0.33, random_state=0)
gnb = GaussianNB()
y_pred = gnb.fit(X_train, y_train).predict(X_test)
```

```
from sklearn import metrics

# Model Accuracy: how often is the classifier correct?
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.8128835071510231
```

4) **Support Vector Machine:**

In the SVM , we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well(i.e by maximizing the distance between the 2 marginal planes so that the points are classified into the both classes with much better accuracy) .
In our case we have used SVM after Naive bayes classifier because SVM is one of the best known classifier as we know in naive bayes based on inductive bias we assume that all the attributes are independent of each other whereas SVM looks at the interactions between them to a certain degree , we found a slight increase in the accuracy as compared to Naive Bayes classifier i.e we found about 0.9% increase in the accuracy about 82.09 as compared to naive bayes where it was about 81.2 of But as we know SVM is not really great when it's used for large dataset (like in our case where we have about 1.3 lakhs entries in our dataset)
Hence we decided to test other classification models where we can expect better accuracy especially the ensemble models

```
In [33]: from sklearn.model_selection import train_test_split
         X_train, X_test, y_train, y_test = train_test_split(df4, df3, test_size=0.3,random_state=109) # 70% training and 30% test

In [34]: #Import svm model
         from sklearn import svm

         #Create a svm Classifier
         clf = svm.SVC(kernel='linear') # Linear Kernel

         #Train the model using the training sets
         clf.fit(X_train, y_train)

         #Predict the response for test dataset
         y_pred = clf.predict(X_test)

In [36]: from sklearn import metrics

         # Model Accuracy: how often is the classifier correct?
         print("Accuracy:",metrics.accuracy_score(y_test, y_pred))

         Accuracy: 0.82086028128289

In [ ]:
```

5.) **Adaboosting :**

It is one of the ensemble based models wherein the predictions of several weak classifiers are combined together to give a better result with a high degree of confidence. This model was chosen because it is highly unlikely to have the problem of overfitting because each of the models have low variance and combined together doesn't lead to overfitting. In boosting all the data points are not given equal weight.This is a sequential technique where the output of the previous model is the input of the successive model. It gives higher weight or priority to the misclassifications i.e the higher weights are assigned to wrong classifications and lower weight to the correct classifications. We made use of Adaboosting after SVM and we saw a significant increase in the accuracy as compared to SVM. The accuracy was about 89.39 i.e around a 7% increase which was a huge improvement.
But one of the disadvantages of this is that it is sensitive to outliers. It keeps giving higher weights to outliers until it is correctly classified which adversely affects the model. Hence we opted for another ensemble model.

```
X = df.drop(['satisfaction'], axis = 1)
y = df['satisfaction']
X_train,X_val,Y_train,Y_val = train_test_split(X,y,test_size=0.25,random_state=28)

adb = AdaBoostClassifier()
adb_model = adb.fit(X_train,Y_train)

adb_model.score(X_val,Y_val)
```

```
0.8939328611025562
```

6.) **Random Forest:**

It is also another type of ensemble model and follows the principle of bagging. Over here multiple weak classifiers classify the different datasets parallely and then are combined together to provide a result. This combining is usually done through voting for classification datasets i.e based on mode of all the classes that were outputted by all the classifiers , Bagging decreases variance and solves over-fitting issues in a model , Hence we made use of this model to expect the best accuracy as there is very less scope of overfitting and got the best accuracy out of all models we had used to classify this dataset , i.e 95.12 which was about 5% increase in the accuracy as compared to the adaboost classifier.

```
In [4]: X = df.drop(['satisfaction'], axis = 1)
        y = df['satisfaction']
        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.33)
        clf = RandomForestClassifier(n_estimators = 100)
        clf.fit(X_train, y_train)

Out[4]: RandomForestClassifier()

In [5]: y_pred = clf.predict(X_test)
        metrics.accuracy_score(y_test, y_pred)

Out[5]: 0.9512143907048366

In [ ]:
```

5) Conclusion ,Contribution of each team member + References + [optional]

Conclusion :

After analyzing many Machine Learning algorithms we came to a conclusion that Random Forest model have better accuracy this model is also not overfitting data The 2 most important factors to be considered are Inflight entertainment and seat Comfort .The airline organisation can try to improve these two features to get more customer satisfaction after these two features it can concentrate on ease of online booking.

Onboard Service is the most important flight service attribute and Online Boarding is the most important pre-flight service attribute for passengers. The flight arrival delay of some less time is positively perceived by passengers but Departure/Arrival Time Convenience is considered a less important attribute by passengers. The most satisfied group of customers are mostly female business passenger's traveling in the business class. On the other hand, the satisfaction can be significantly improved in a large group of business customers traveling in eco class.

SNo	Model	Accuracy
1	KNN model	70.8
2	Logistic Regression	74.02
3	Naïve Bayes model	81.2
4	SVM model	82.08
5	Adaboost	89.39
6	Random Forest	95.12

Contributions:

Y Nikhil Bhardwaj : Support Vector Machine , Naive Bayes Classification

P Sai Varshith : Logistic Regression, Adaboosting

P J Subramanya Hande : K- Nearest Neighbors, Random Forests.

Preprocessing : Equal Contributions from all

References :

1. https://www.academia.edu/download/55007446/Article_313.pdf
2. https://www.sciencedirect.com/science/article/pii/S0969699718304873
3. https://www.inderscienceonline.com/doi/abs/10.1504/IJBIR.2017.082829
4. https://www.tandfonline.com/doi/abs/10.2753/MTP1069-6679190407
5. https://www.kaggle.com/sjleshrac/airlines-customer-satisfaction