CUSTOMER GRATIFICATION IN AIRLINE INDUSTRY

Ms. Gavoury R[1], Sujitha D[2], Tharani R[3]

[1]M.Sc. Information Technology, M.O.P. Vaishnav College for Women, India
E-mail: gavouryr.csc@mopvaishnav.ac.in
[2]M.Sc. Information Technology, M.O.P. Vaishnav College for Women, India
E-mail: sujithadmy@gmail.com
[3]M.Sc. Information Technology, M.O.P. Vaishnav College for Women, India
E-mail: tharaniravi2702@gmail.com

ABSTRACT:

The airline industry is always focused on providing excellent customer service. It is crucial to determine the specific levels of satisfaction with which the airline passenger is satisfied or dissatisfied. The primary focus of this paper is on addressing the imperative need to enhance passenger satisfaction in the airline industry. It's crucial for the industry's financial performance, customer retention, and competitive positioning. The research is aimed at analysing the competition within the aviation industry and identifying the key factors contributing to its success. This study presents the process of constructing Machine Learning models, starting with Data Acquisition, Data Cleaning, Exploratory Data Analysis, Preprocessing, and Model Building. This paper contains various methods and techniques that are used to determine the most accurate model and various visualization tools to understand the data. This paper also helps to understand the importance of features in predicting passenger satisfaction and to identify the correlation between features. Through the implementation of machine learning classification algorithms such as Support Vector Machine, Logistic Regression, Gaussian NB, Decision Trees, and Random Forest, it is observed that the Random Forest Classifier achieves the highest accuracy rate of 96%. Hence, it will help airline companies to adjust their service value and demand to satisfy customers' satisfaction.

Keywords: Machine Learning, Classification, Random Forest

1. INTRODUCTION

The aviation industry has faced significant challenges recently, leading to financial losses and bankruptcy filings for several airlines worldwide. However, as travel gradually resumes, there's an expectation of increased demand. This study delves into the competitive landscape of the aviation sector and identifies key factors crucial for its recovery. Using various classification models, such as SVM, Logistic Regression, Gaussian NB, Decision Trees, and Random Forest, the research aims to predict customer satisfaction levels. Notably, the Random Forest Algorithm yields an impressive accuracy rate, with Inflight Wi-Fi Service identified as a significant determinant of customer satisfaction.

Revitalizing the airline industry amidst economic downturns necessitates enhancing service offerings to restore passenger confidence. Leveraging advancements in technology, particularly Machine Learning, airlines can discern pivotal aspects influencing passenger satisfaction and categorize customer experiences effectively. This study presents a comprehensive framework for developing Machine Learning models, encompassing Data Acquisition, Data Cleaning, Exploratory Data Analysis, Preprocessing, and Model Building stages. Results indicate that the Random Forest algorithm outperforms others, boasting high accuracy while requiring a shorter modeling period.

Air transportation stands as an indispensable facet of modern life, facilitating global connectivity and accessibility. This discourse underscores the significance of full-service airlines with online-based amenities in meeting passenger needs. Past research underscores the pivotal role of available facilities and services in shaping passenger satisfaction.

2. LITERATURE REVIEW

2.1 Machine Learning:

"Pattern Recognition and Machine Learning" by Christopher M. Bishop: This comprehensive textbook provides a thorough introduction to the principles of pattern recognition and machine learning. It covers various topics including supervised and unsupervised learning, neural networks, and probabilistic graphical models.

2.2 Classification Algorithms:

"Introduction to Statistical Learning: with Applications in R" by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani provides a practical introduction to a variety of statistical learning methods. These methods encompass classification algorithms such as logistic regression, decision trees, and support vector machines. It offers insights into theoretical concepts alongside practical implementations.

"Python Machine Learning" by Sebastian Raschka and Vahid Mirjalili covers essential machine learning concepts and techniques using Python. It includes chapters on classification algorithms, delving into topics such as k-nearest neighbors, decision trees, and ensemble methods.

2.2 Random Forest Algorithm:

"Random Forests" by Leo Breiman: This seminal paper introduces the Random Forest algorithm, which is an ensemble learning method for classification and regression tasks. It discusses the underlying principles, construction of decision trees, and the aggregation process that forms the Random Forest model.

"Random Forests" by Trevor Hastie and Robert Tibshirani: This book chapter provides a detailed overview of Random Forests, focusing on the algorithm's strengths and weaknesses. It discusses

practical considerations for building and tuning Random Forest models and compares them with other machine learning approaches.

3. METHODOLOGY

The notebook used in making the model uses the help of Jupyter Notebook. The workflow of this research is presented in Figure 1. The flow of the research was carried out by retrieving data from the web page of the dataset provider, namely Kaggle; after the data is obtained, data cleaning will be carried out, which includes checking whether there is data that is not balanced and checking whether there is empty data; then the Exploratory Data Analysis process is carried out where the search for essential points that can be visualized from the dataset will be carried out; after the understanding of the data is complete, the Pre-processing stage will be carried out where there is a Label Encoding and Outlier Removal process; Then the last one will be making a model with a predetermined configuration.

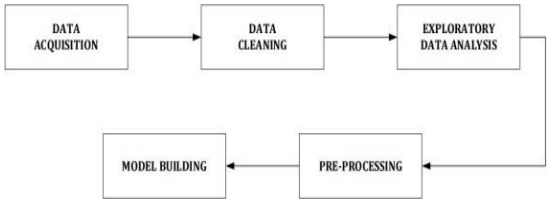


Figure 1. Research Methodology

3.1 DATA ACQUISTION

The data acquisition stage aims to download data from trusted sources before the data is stored, processed, pre-processed, and used for other purposes. This process begins with retrieving relevant information, changing the data as needed, and calling the dataset into the notebook. In this study, the dataset used came from Kaggle, namely "Airline Passenger Satisfaction". This dataset contains passenger survey data from an airline. The survey covers passenger numbers to passenger satisfaction. This dataset contains 25 columns and 103904 rows of data. Details of the dataset can be seen in Figure 2.

	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	...	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes	satisfaction
0	70172	Male	Loyal Customer	13	Personal Travel	Eco Plus	460	3	4	3	...	5	4	3	4	4	5	5	25	18.0	neutral or dissatisfied
1	5047	Male	disloyal Customer	25	Business travel	Business	235	3	2	3	...	1	1	5	3	1	4	1	1	6.0	neutral or dissatisfied
2	110028	Female	Loyal Customer	26	Business travel	Business	1142	2	2	2	...	5	4	3	4	4	4	5	0	0.0	satisfied
3	24026	Female	Loyal Customer	25	Business travel	Business	562	2	5	5	...	2	2	5	3	1	4	2	11	9.0	neutral or dissatisfied
4	119299	Male	Loyal Customer	61	Business travel	Business	214	3	3	3	...	3	3	4	4	3	3	3	0	0.0	satisfied

Figure 2. Dataset Detail

3.2 DATA CLEANING

The data cleaning stage prepares data for analysis by removing irrelevant or inappropriate data. The data in question has a negative impact on the model or algorithm to be made. Data cleaning is not only to dispose of data but can also be interpreted as a step to improve data. In this study, data cleaning is done by checking the null value.

Table 1. Data Cleaning Process

Column	**After Data Cleaning**
Arrival_Delay_in_Minutes	Filled with mean value
Gender	Filled with categorical data (eg.0/1)
Customer_Type	
Type_of_Travel	
Class	

3.3 EXPLORATORY DATA ANALYSIS

The EDA stage is an essential process for conducting an initial investigation of the data used to find patterns and anomalies and test hypotheses with the help of statistical visualization. In this study, EDA was conducted to visualize data such as the number of satisfied or dissatisfied passengers. An example of the EDA stages can be seen in Figure 3.

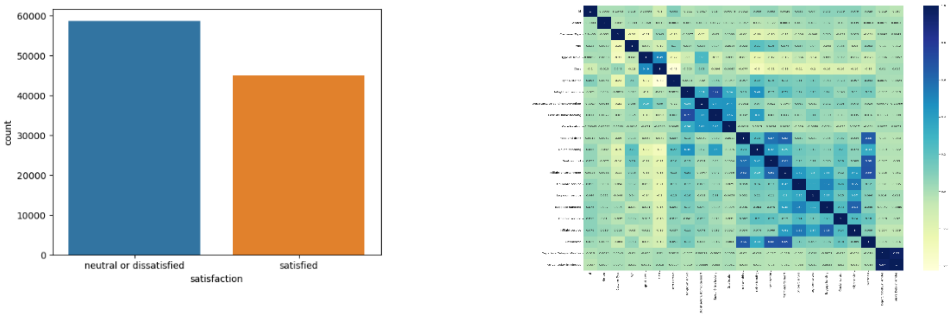


Figure 3. Exploratory Data Analysis

3.4 PRE – PROCESSING

The preprocessing stage involves transforming or encoding data to make it easily understandable for machine learning algorithms. Its primary objective is to create an accurate and predictable model. In this study, pre-processing uses Label Encoder to encode the data and remove outliers contained in the data. Details of the data encoding process and the outlier removal process can be seen in Figure 4.

```
#PREPROCESSING
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
a['Gender']=le.fit_transform(a['Gender'])
a['Customer Type']=le.fit_transform(a['Customer Type'])
a['Type of Travel']=le.fit_transform(a['Type of Travel'])
a['Class']=le.fit_transform(a['Class'])
```

id	Gender	Customer Type	Age	Type of Travel	Class
70172	1	0	13	1	2
5047	1	1	25	0	0
110028	0	0	26	0	0
24026	0	0	25	0	0
119299	1	0	61	0	0

Figure 4. Preprocessing

4. RESULT AND DISCUSSION

4.1 MODEL BUILDING

4.1.1 SUPPORT VECTOR MACHINE

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is 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.

4.1.2 DECISION TREE CLASSIFIER

Decision Tree is a tree structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision Tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

4.1.3 NAIVE BAYES

Naive Bayes is a simple and effective classification algorithm based on Bayes' theorem with the assumption of feature independence. It is a machine learning algorithm commonly used for classification tasks, especially in natural language processing (NLP) and text mining. It calculates the probability that a data point belongs to a certain class given its features. It is efficient and easy to implement, making it a popular choice for classification problems, particularly with limited training data.

4.1.4 RANDOM FOREST CLASSIFIER

Random Forest is a machine learning algorithm that belongs to the supervised learning technique. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

4.1.5 LOGISTIC REGRESSION

Data visualization is the graphical representation of data to help people understand and interpret it more easily. It involves creating visual representations of data sets to convey insights, patterns, and trends that might be less apparent when examining the data in its raw, numerical form. Data visualization is a crucial component of data analysis and data communication, as it enables individuals to make data-driven decisions, identify patterns, and communicate complex information effectively.

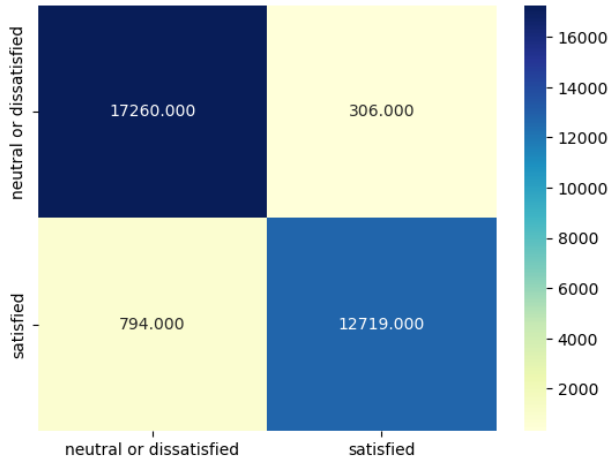
	Support Vector Machine	Decision Tree Classifier	Naive Bayes	Random Forest Classifier	Logistic Regression
Model	LinearSVC(random_state=0)	DecisionTreeClassifier()	GaussianNB()	(DecisionTreeClassifier(max_features='sqrt', r...	LogisticRegression()
Accuracy	0.780205	0.943274	0.811577	0.963223	0.716593

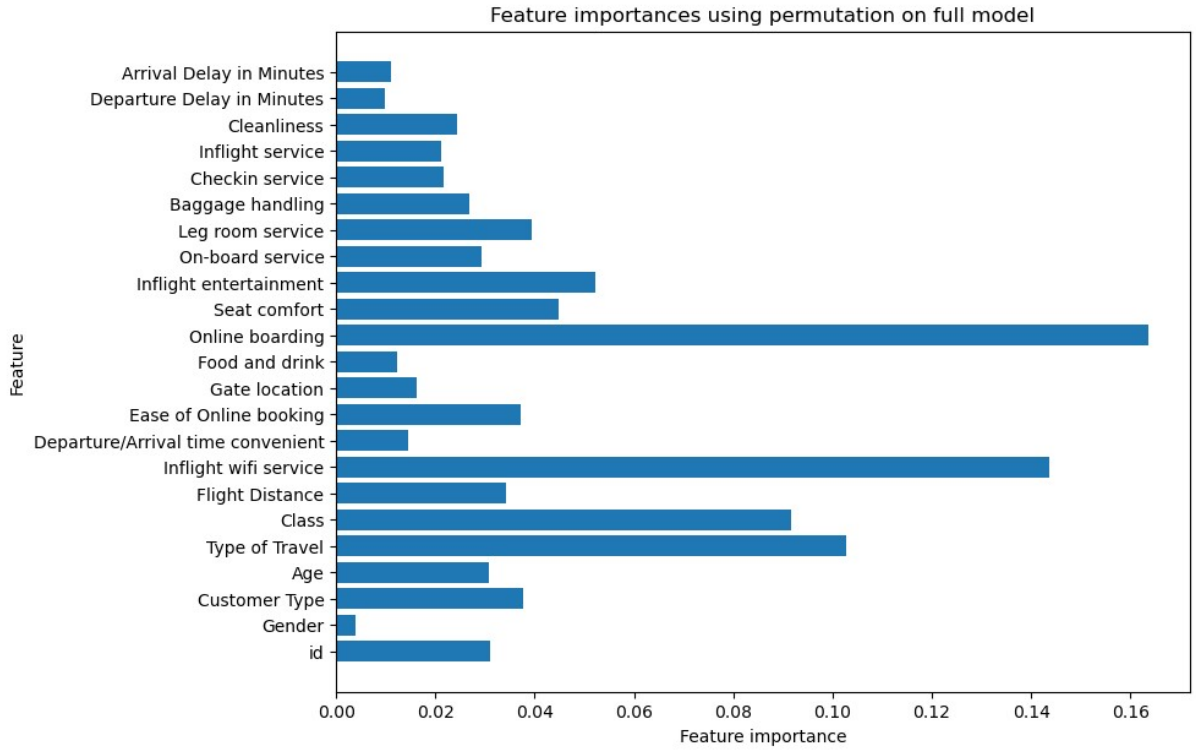
Therefore, Random Forest Classifier gives highest accuracy with 96%.

Figure 5. Accuracy Score

5. CONCLUSION

This study will use five algorithm models to carry out the classification process: Random Forest, Logistic Regression, Support Vector Machine, Decision Tree Classifier and Naive Bayes. The Random Forest algorithm, with an accuracy value of 0.963223 (96.3%).






```
prediction = random_forest.predict(np.array([[5676,0,0,67,1,2,1,287,3,4,4,3,4,3,2,4,5,4,3,8,9,7,5]]))
print(prediction)
```

```
['satisfied']
```

```
C:\ProgramData\anaconda3\lib\site-packages\sklearn\base.py:420: UserWarning: X does not have valid feature names, but RandomFor
estClassifier was fitted with feature names
  warnings.warn(
```

```
prediction = random_forest.predict(np.array([[5047,1,1,25,0,0,235,3,2,3,1,1,1,5,3,1,4,1,1,2,7,9,9]]))
print(prediction)
```

```
['neutral or dissatisfied']
```

```
C:\ProgramData\anaconda3\lib\site-packages\sklearn\base.py:420: UserWarning: X does not have valid feature names, but RandomFor
estClassifier was fitted with feature names
  warnings.warn(
```

Figure 6. Result of Random Forest

With the influence of technology in all industrial fields, airlines can now use Machine Learning to find the essential points that can make passengers feel satisfied with airline services. The airline can also classify the rating given by the passenger to find out whether the passenger is satisfied or not with the service that has been provided. It can be concluded that the best model in this case study is Random Forest with a Acuracy value of 96% and the model generation time is the fastest compared to other algorithms that have been made in this study.

6. BIBLIOGRAPHY

[1] K. Hulliyah, "Predicting Airline Passenger Satisfaction with Classification Algorithms," IJIIS: International Journal of Informatics and Information Systems, vol. 4, no. 1, pp. 82–94, Mar.2021.

[2] J. Choi, H. Seol, S. Lee, H. Cho, and Y. Park, "Customer satisfaction factors of mobile commerce in Korea," Internet Res., vol. 18, no. 3, pp. 313–335, 2008.

[3] S. Chen, B. Shen, X. Wang, and S.-J. Yoo, "A Strong Machine Learning Classifier and Decision Stumps Based Hybrid AdaBoost Classification Algorithm for Cognitive Radios," Sensors, vol. 19, no. 23, p. 5077, Nov. 2019, doi: 10.3390/s19235077.

[4] Huang, Y., & Crotts, J. C. (2019). Understanding passenger satisfaction with airline service quality: Evidence from the global airline industry. Journal of Air Transport Management, 74, 52-63.

[5] Kwon, O., & Park, S. (2019). Airline service quality, perceived value, and passenger satisfaction: Evidence from South Korea. Journal of Air Transport Management, 75, 27-34.

[6] Xie, W., Peng, G., & Zhang, Y. (2021). Predicting Passenger Satisfaction in the Airline Industry Using Machine Learning Techniques. Journal of Air Transport Management, 93, 101986.

[7] Li, S., Zhang, A., & Wang, Y. (2019). Predicting Passenger Satisfaction in the Airline Industry: A Comparison Study of Machine Learning Methods. Journal of Advanced Transportation, 2019, 7136958.

[8] Tussyadiah, I. P., & Park, S. (2018). Consumer Evaluation of Online Information in Travel Decision-Making: The Case of TripAdvisor. In Information and Communication Technologies in Tourism 2018 (pp. 165-177). Springer, Cham.

[9] Chen, P., & Xie, X. (2020). Passenger Satisfaction with Airline Services: A Comprehensive Review and Future Research Directions. Journal of Air Transport Management, 84, 101768.

[10] Park, J. H., & Kim, S. (2019). A Study on the Factors Influencing Passenger Satisfaction and Loyalty in the Airline Industry: Focusing on Service Quality and In-Flight Experiences. Sustainability, 11(14), 3973.