1. ABSTRACT

The objective of this case study is to predict the satisfaction level of Airlines Customer using exploratory data analysis and machine learning algorithms and gain insights to increase customer base. With 100,000+ observations and 25+ attributes, the data dataset used is quite comprehensive and reliable. Most features considered are related to customer survey data and customer demographics. In this project, we will perform exploratory analysis and classification regression analysis using machine learning algorithms to find the underlying pattern and make some insights towards our objective. Models are comparisons based on their predictive power as well as their interpretability. Finally, features which best help with the classification are drawn from the data to narrow the focus upon those factors that contribute to customer satisfaction the most.

2. INTRODUCTION

A vast majority of the economic model of airline companies is based upon their customer base. This dependence goes beyond just the direct fare revenue generated. Most of the profits this industry sees are in terms of the value of Air Miles travelled which relates to how well they retain/ expand their customer base. Like most services, customer satisfaction plays a significant role in the aviation industry. In this project, we will explore an airlines' customer satisfaction dataset to answer the questions: which factors contribute to customer satisfaction the most and how to maximize it.

3. OVERVIEW OF THE DATA

The data source for the following project has been taken from the open-source repository: Kaggle for which the link can be found here. Broadly speaking, the dataset contains entries relating to customer demographics and customer preferences. It consists of 100,000+ observations and 20+ attributes. Each record is categorized by respective customer IDs. For each observation, the following factors have been recorded:

```
Gender
Customer Type
Age
Type of Travel
Class
Flight Distance
Inflight wifi service
Departure/Arrival time convenient
Ease of Online booking
Gate location
Food and drink
Online boarding
Seat comfort
Inflight entertainment
On-board service
Leg room service
Baggage handling
Checkin service
Inflight service
Cleanliness
Departure Delay in Minutes
Arrival Delay in Minutes
satisfaction
```

This report will further investigate which attributes are more closely related to customer satisfaction (Target) and by how much, thereby enabling accurate future predictions possible.

3.1 STRUCTURE OF THE DATA & MISSING VALUES

Before any further analysis can be conducted, it is necessary to ensure the reliability of the dataset. Checking for missing data, a total of 310 values for “Arrival Delay” were observed. Therefore, these observations were dropped from the dataset. This is not expected to have a significant impact on the analysis due to only a very small count of omitted values.

Further, it is important to note the “type” for each attribute at this point as any kind of analysis may give misleading outcomes for mistaken data types. Thus, the dataset is transformed to represent the correct attribute type and the list is as follows:

Categorical Features	Ordinal Features	Numerical Features	Target Feature
Gender Customer Type Type of Travel Class	Inflight Wi-Fi Service Departure/Arrival Time Convenient Ease of Online Booking Gate Location Food and Drink Online Boarding Seat Comfort Inflight Entertainment On-Board Service Leg Room Service Baggage Handling Check in Service Inflight Service Cleanliness	Age Flight Distance Departure Delay in Minutes Arrival Delay in Minutes	Satisfaction

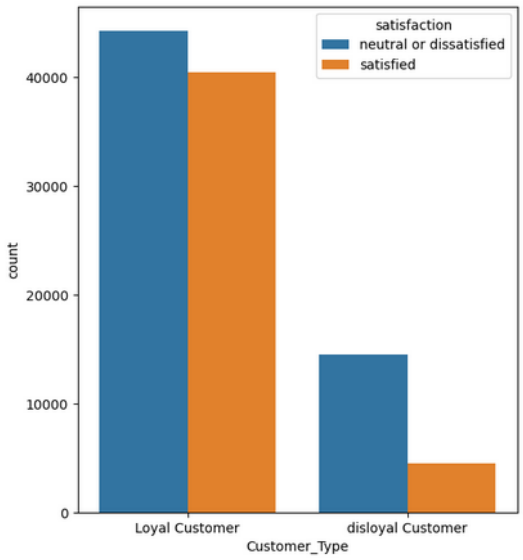
3.2 NORMALIZATION

For the purpose of running models like Random Forest and Logistic Regression etc., Feature scaling for continuous variables is implemented to curb the “pull” effect caused by attributes with larger values.

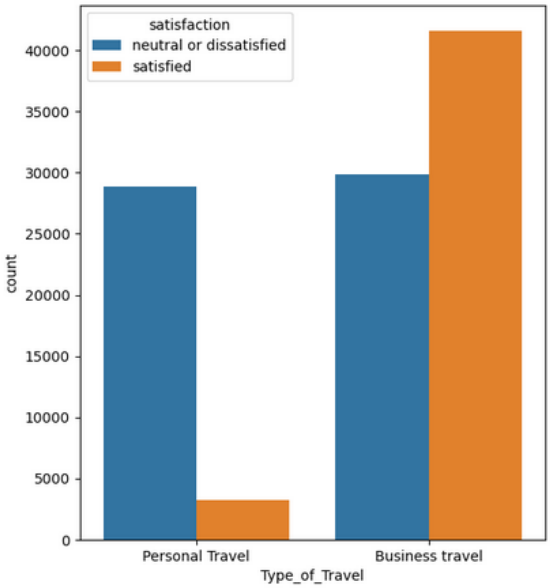
3.3 UNIVARIATE ANALYSIS

Now that the data has been somewhat cleaned and preprocessed, each attribute can be analyzed independently. The focus is to identify any trends from basic plots and statistical measures to determine any underlying patterns. This might even help pinpoint the variables most significantly related to customer satisfaction.

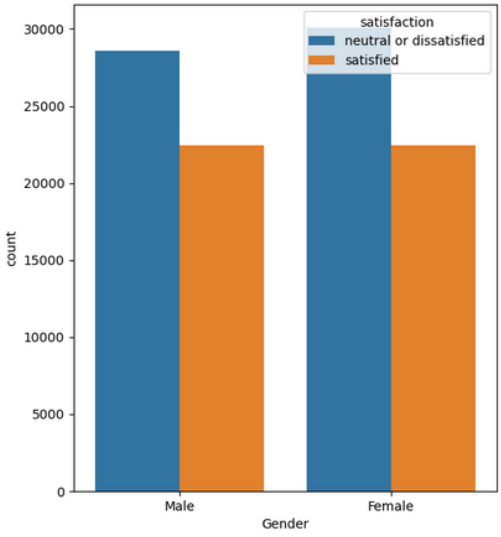
ATTRIBUTES: CUSTOMER TYPE, TYPE OF TRAVEL, GENDER & CLASS



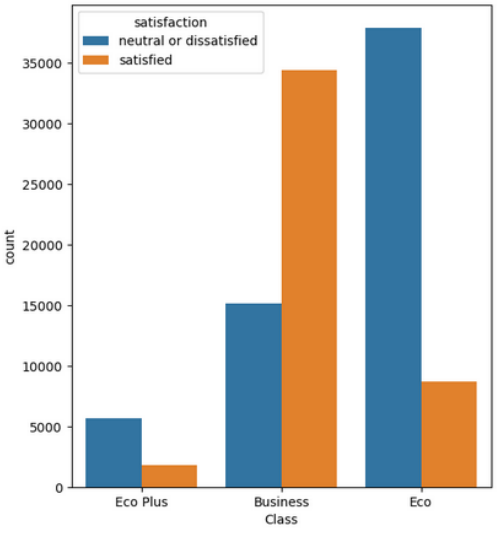
Satisfaction based upon the Customer Type



Satisfaction based upon the Type of Travel



Satisfaction based upon Gender



Satisfaction based upon the Travel Class

CUSTOMER TYPE

From the summary statistics and the plot, we can observe most of the customers are classified as Loyal Customers. Also, Loyal Customers have the same proportion of satisfied and unsatisfied customers whereas most Disloyal Customers are unsatisfied. Thus, Customer Type seems to play some role in determining whether a customer is satisfied or not.

GENDER

The male to female ratio is almost the same. Also, the percentage of unsatisfied customers is slightly higher than the satisfied customers, but the ratio is the same across both groups. Thus, we cannot clearly state whether gender is an important factor in determining customer satisfaction and expect the same from our models.

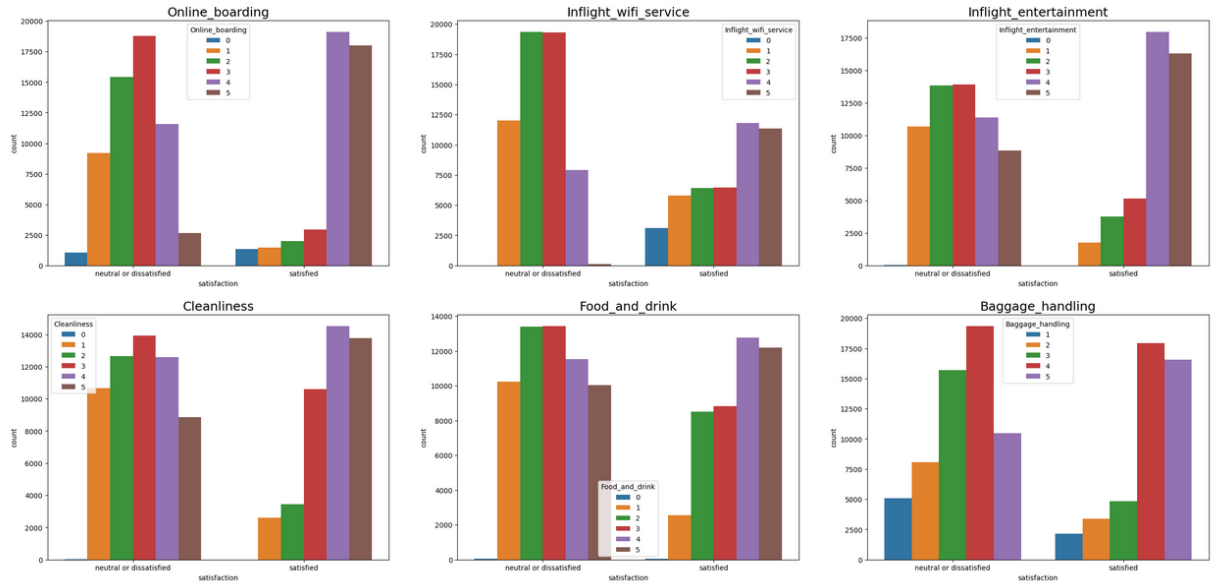
TYPE OF TRAVEL

Business travel holds a larger share. Also, Business travels tend to be largely satisfied whereas personal travels do not. This is a clear difference and is expected to play some role in determining the outcome.

CLASS

From the summary statistics and the plot, we can observe that only Business Class has a positive ratio for satisfaction. This is a much-expected phenomenon as better services are provided for business class and satisfaction, in turn, is high. Thus, Class might not be the best attribute for modelling.

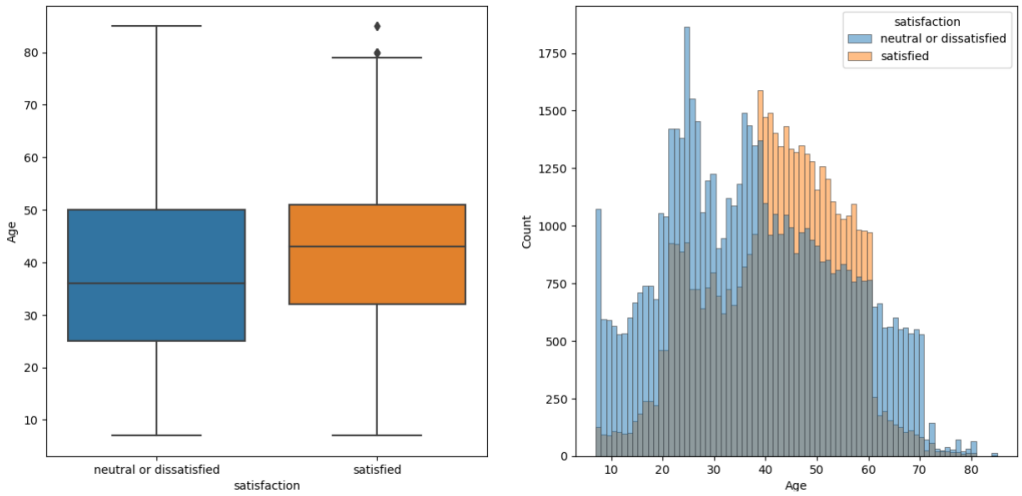
ORDINAL ATTRIBUTES



A general trend of higher scores for satisfied customers and a normally distributed trend for unsatisfied customers can be observed.

AGE

The distribution of satisfied vs unsatisfied based upon the Class is plotted below.

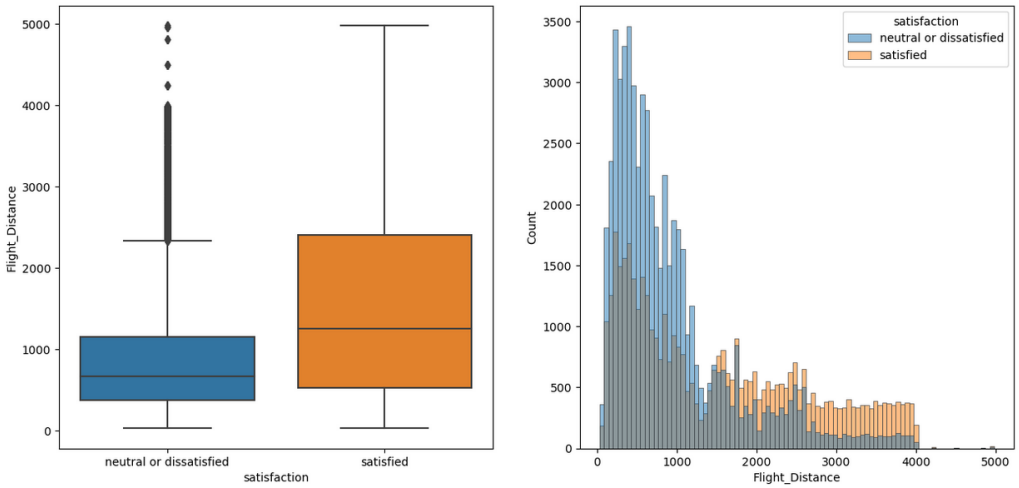


Based upon the statistical observations, the mean age of the sample is around 40 years with 50% of the passengers in the range of 27 to 51 years. This usually represents the working are group and the data

seems to be consistent with real world expectations. Also, the distribution is similar across both the groups. Thus, we expect age to not to play a significant role in the models or have a minimal effect.

FLIGHT DISTANCE

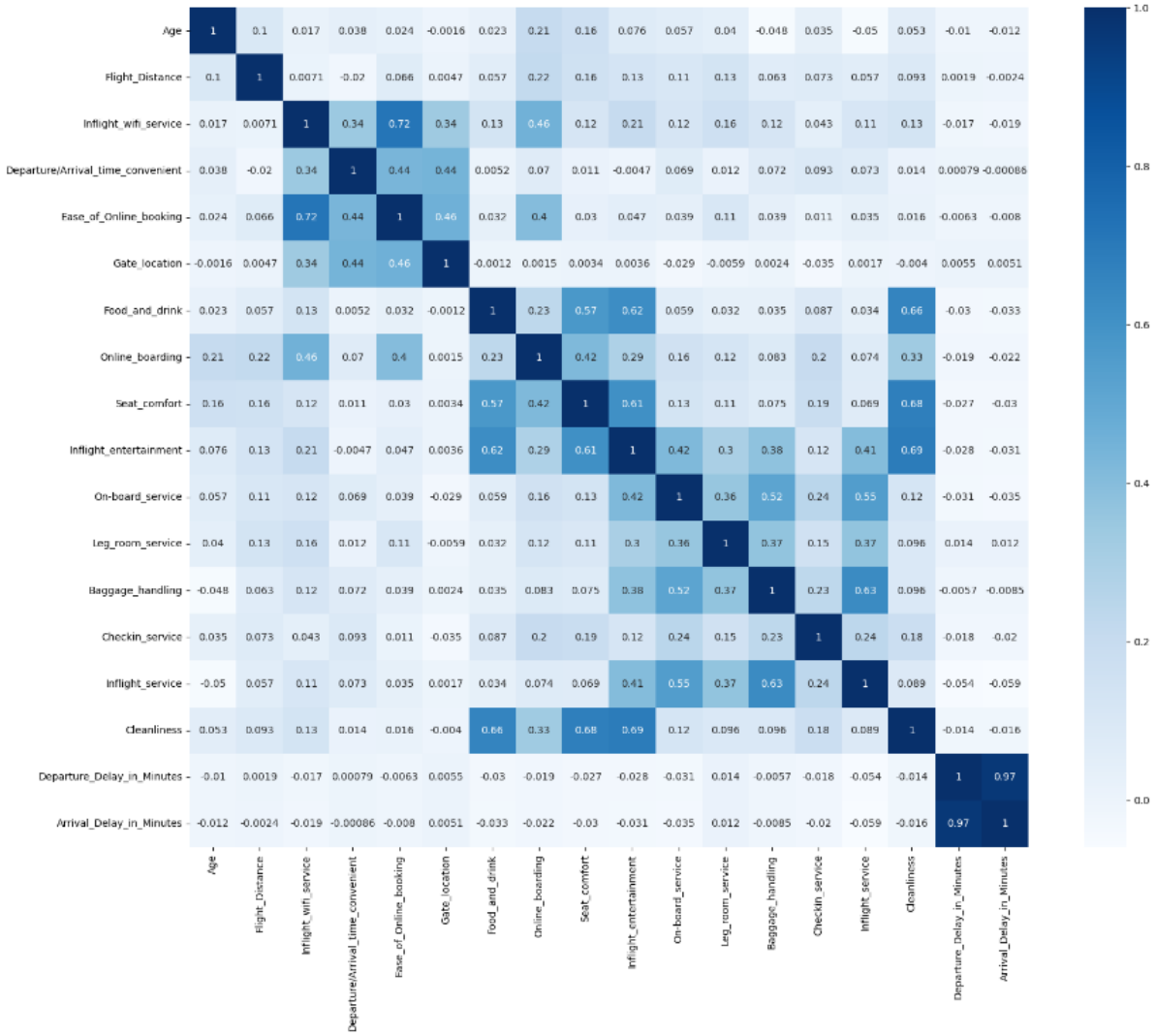
The distribution of satisfied vs unsatisfied based upon the Flight Distance is plotted below.



The trend seems to follow a Poisson distribution which approximates to a normal distribution for large datasets under regression. This is an expected feature of arrival distribution. No significant anomaly, except for some outliers, can be observed and hence, we expect Flight Distance to not play a significant role in our classification models.

3.4 MULTIVARIATE ANALYSIS: CORRELATION AMONG ATTRIBUTES

Another major factor that may affect the classification models is the autocorrelation existing between two variables that may yield misleading results. If a variable does not convey any additional information, it should be dropped. For the purposes of this project, we will consider correlation between variables and set a threshold of 70%. Based on the below matrix, some variables are dropped.



Dropped attributes: Arrival Delay in minutes, Ease of Online Booking, Cleanliness

3.5 KEY FINDINGS AND TAKEAWAYS

- The dataset is quite balanced in terms of the gender of the sample with both having the same proportion of satisfaction level.
- Most Loyal Customers and Business Travel Customers tend to be satisfied. The Business Class seems to do well for this sample.
- Satisfied customers generally tend to rate the services provided at 4/5 or 5/5 while the distribution of rating for unsatisfied customers is generally normally distributed.
- Arrival Delay and Departure Delay seem to be correlated. This is expected as the departure delay also causes the flight to be late. Therefore, only Departure Delay will be considered in future regression modeling.

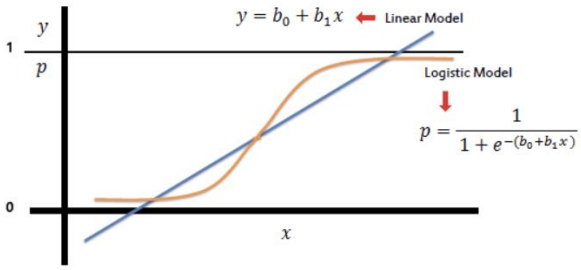
4. CLASSIFICATION MODELS

In this section, various classification models are estimated based on the cleaned dataset from the above steps. The purpose of this section is to:

1. Determine the parameters that affect customer satisfaction the most.
2. Validate the observations from exploratory analysis.
3. Test out various models and suggest one with the best accuracy score, performance and interpretability.

4.1 LOGISTIC REGRESSION

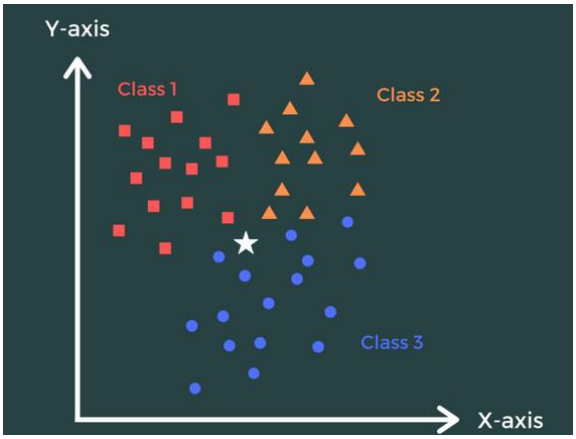
First and foremost, a logistic model was considered since this model can handle both categorical and numerical variables and gives a binary outcome based on probability. An advantage of this model is the interpretability of the regression model.



Binary Classification using Logistic Regression

4.2 K NEAREST NEIGHBORS

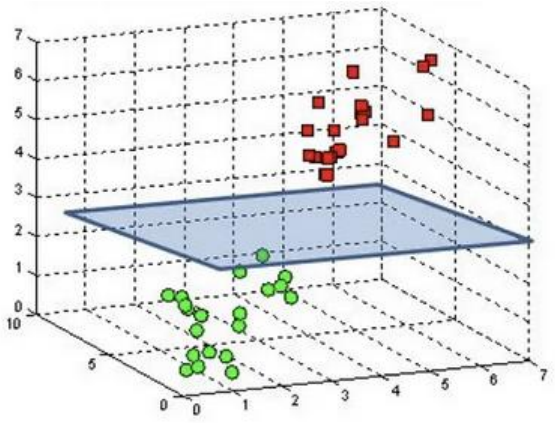
This model classifies each datapoint based on its proximity to other datapoints. With the implementation of Hyperparameter Tuning and Cross Validation, better results were obtained in terms of the accuracy of the model. One major drawback of the model is the non-interpretability of the model.



Representation of KNN classification

4.3 SUPPORT VECTOR MACHINE

This model classifies each datapoint based on a distinction drawn between each class as a hyperplane. Hyperparameter Tuning and Cross Validation were initially tried but were not implemented due to the algorithm being very computationally heavy. better results were obtained in terms of the accuracy of the model. One major drawback of the model is the non-interpretability of the model.



Representation of SVM Classifier

4.4 RANDOM FOREST

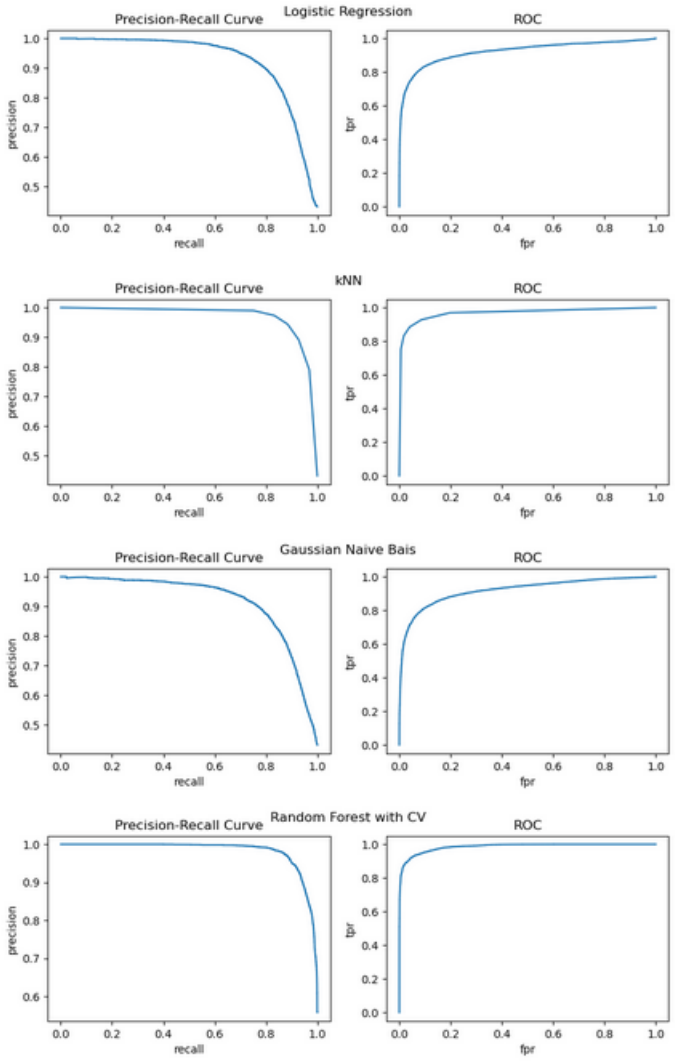
This model creates a hierarchical tree based on the Information Gain of each attribute. Splits are created on maximum IG with the objective of obtaining the cleanest splits (least misclassification). Hyperparameter Tuning and Cross Validation were performed to limit the number of splits and avoid overfitting. Better results were obtained in terms of the accuracy of the model. With The help of feature importance, the most relevant features that explain Satisfaction can be predicted.

4.5 MODEL COMPARISON

Model	Accuracy	Pros/Cons List
Logistic Regression	87%	Interpretation of the model is simple. Relatively Lower accuracy.
K-Nearest Neighbors	93%	Hardware heavy & long time to execute. Interpretation of results is hard.
Support Vector Machine	93%	Hardware heavy & long time to execute. Interpretation of results is hard.
Naïve Bayes	86%	Low accuracy
Random Forest	94%	Computationally heavy. Can help determine most important attributes based on their Information Gain

Logistic regression, though easy to interpret, has a comparatively low accuracy score along with Naïve Bayes. K-Nearest Neighbors has a good accuracy score along with cross validation and less overfitting, but the result is not so interpretable. Support Vector Machine suffers from the same drawback. Random forest seems to perform well across all fields and might be a suitable model for our purposes. Conclusion can only be drawn after testing for False Positives and True Negatives.

4.6 AUC AND ROC ANALYSIS



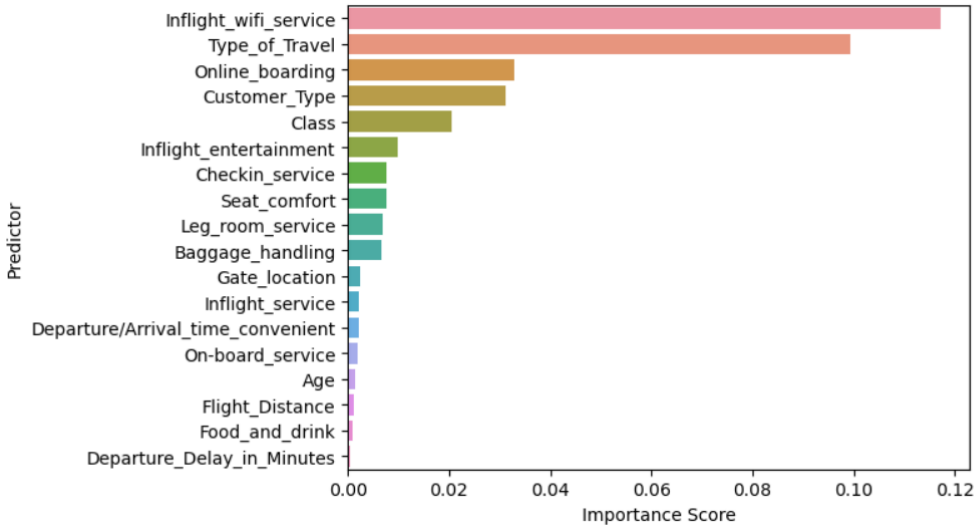
AUC ROC Plot for Classification Models

In order to address the False Positives and True Negatives, Accuracy score may not be independently sufficient to rate the performance of the model. Hence AUC and ROC curves are also considered. Since Random Forest produces the most area under the curve, it seems to have the best classification performance.

5. RESULTS AND RECOMMENDATIONS

5.1 SELECTED MODEL

Since Random Forest has an accuracy of 94% along with the best area under curve performance for AUC and ROC, it can be concluded that it has the best performance measures. This, coupled with the fact that it also addresses overfitting by Hyperparameter Tuning and Cross Validation, the results seem to be quite reliable. The most important features of this model based on the feature score are as follows:



5.2 MOST SIGNIFICANT FEATURES

According to our selected model, the most important features that contribute to customer satisfaction are Inflight WIFI service, the Type of Travel, the option for Online Boarding, Customer Type and Class of travel. Interestingly, features like Departure delay and Flight Distance seem to have very less effect on satisfaction of the customers. Moreover, these results seem to be in line with the exploratory analysis performed and validate our earlier observations.

5.3 MAXIMIZING SATISFACTION

From all the features analyzed, there seems to be a trend that the overall customers experience is more governed by the services provided in flight. Therefore, the airlines should focus and improve the same.

Also, there seems to be considerable parity in customer satisfaction in business and economy class, therefore, the flights can incorporate a larger business section and bring economy class services to par.

5.4 IMROVEMENTS

The dataset seems to consist of only the survey conducted by the airlines. This method, though effective, may not be able to capture the entire picture of how the customers feel. A better method would be to actively monitor customer responses at each stage of their flight experience such as collective usage data of in-flight Wi-Fi and entertainment services.

6. REFERENCES

Logistic Regression: https://www.ibm.com/topics/logistic-regression

Regression Code: https://www.saedsayad.com/logistic_regression.htm

Random Forest Feature Selection: https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

ROC & AUC: https://www.statology.org/plot-roc-curve-python/

Dataset: https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

Machine Learning Algorithm: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47