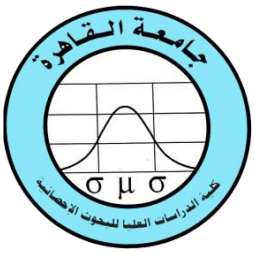

Machine Learning and Customer Behavior: Predicting Purchasing Patterns with Xgboost

Term Paper

Submitted by

Asya Ahmed El- Sayed

Supervised By

Prof. Dr. Mohamed Abd El-Hamid

2022-2023

Acknowledgment

We would like to express our sincere appreciation to Professor Mohamed Abd El-Hamid for his invaluable guidance and supervision during this project. His expertise, patience, and constructive feedback were invaluable in helping us to refine our research questions and analysis methods.

We would also like to extend our gratitude to British Airways for providing the dataset used in this research. The insights gained from this dataset have been invaluable in helping us to understand customer behavior and develop a predictive model to identify customers likely to book holidays with British Airways.

Finally, we would like to thank all members of the data science department, including professors, teaching assistants, and employees, who have contributed to our academic and professional development. Your knowledge, guidance, and support have been instrumental in helping us to grow as data scientists. Once again, thank you to everyone who has helped us complete this research project.

Abstract

Customer behavior is a crucial factor for any business to understand to maximize profits. The airline industry is highly competitive, and it is essential to understand customer behavior to retain customers and attract new ones. This paper presents a machine learning-based approach to predicting customer behavior and identifying the customers who are most likely to book a holiday with British Airways. The data used in this study consists of historical customer booking data. We used three models XGBoost, CatBoost, and Logistic regression as a baseline model. The data preprocessing techniques included robust scaling, and ADASYN (Adaptive Synthetic)to handle outliers and imbalances in the data.

The tuned XGBoost model outperformed the other models and achieved an accuracy of 91%. The precision of the model for the booking customer class was 95%, while the recall was 87%. These results demonstrate that our model can accurately predict customer booking behavior, and it has the potential to be used to improve customer acquisition and retention strategies in the airline industry.

Table of contents

List of figures

List of tables

1. Chapter One: Introduction

1.1 Introduction

In today's digital age, companies are collecting an enormous amount of data on their customers' behaviors and preferences. Analyzing this data can provide businesses with invaluable insights into their customers, including their buying habits and preferences. Machine learning, a subset of artificial intelligence, is a powerful tool that can be used to analyze customer data and predict future buying behaviors.

This study aims to explore the application of machine learning in predicting customer buying behavior. The paper will focus on understanding the key factors that influence customer buying behavior, such as previous purchase history, and economic factors. By analyzing historical customer booking data, we aim to develop a predictive model that accurately predicts the likelihood of a customer booking a holiday.

The development of a predictive model has significant practical implications for businesses, as it can help them personalize customer experiences, improve marketing strategies, and increase revenue. By identifying the customers who are most likely to make a purchase, businesses can target their marketing efforts more effectively and allocate their resources more efficiently.

The research will involve the analysis of the available data, including customer booking history, and other relevant factors. The paper will explore various machine learning techniques that can be used to develop a predictive model, such as XGboost, Catboost, and logistic regression.

Overall, the goal of this project is to contribute to the understanding of customer buying behavior and provide insights that can help businesses improve their operations. The potential benefits of using machine learning to predict customer buying behavior are significant, and the insights gained from this research can be applied across a wide range of industries not only in the airline industry.

1.2 Motivation

The ability to predict customer buying behavior is a critical success factor for any business, as it enables them to provide personalized customer experiences, improve marketing strategies, and increase revenue. Machine learning has emerged as a powerful tool for analyzing customer data and predicting future buying behaviors. Applying machine learning to customer data has led to significant improvements in business operations, with many businesses now leveraging machine learning models to drive growth and profitability.

This term paper is motivated by the need to explore the application of machine learning in predicting customer buying behavior. This project aims to develop a predictive model that accurately predicts the likelihood of a customer making a purchase, based on their previous purchase history, and other relevant factors.

The potential benefits of using machine learning to predict customer buying behavior are significant. By accurately predicting which customers are most likely to make a purchase, businesses can target their marketing efforts more effectively, personalize customer experiences, and allocate resources more efficiently. This can lead to increased revenue, improved customer satisfaction, and a competitive advantage in the market.

The research in this paper has the potential to contribute to the broader understanding of customer buying behavior and the application of machine learning in the field of marketing and sales. By exploring the various machine learning techniques and data sources that can be used to develop a predictive model, this research can provide valuable insights and practical guidance for businesses looking to leverage machine learning to drive growth and profitability.

1.3 Problem definition

The airline industry is highly competitive, with airlines vying for customers' attention and loyalty. As such, airlines are constantly seeking ways to improve their marketing strategies and drive revenue growth. One way to achieve this is by accurately predicting which customers are most likely to book a holiday with the airline.

The problem addressed in this term paper is to develop a predictive model that can accurately identify customers who are most likely to book a holiday with British Airways. The objective is to use historical customer booking data to build a machine-learning model that can predict the likelihood of a customer booking a holiday with the airline.

To achieve this goal, the project will focus on analyzing a variety of customer data, including travel history, and other relevant factors that may influence a customer's decision to book a holiday with the airline. The predictive model will be developed using various machine-learning techniques, such as XGboost, Catboost, and logistic regression.

The successful development of a predictive model has significant practical implications for British Airways and the airline industry as a whole. By accurately identifying customers who are most likely to book a holiday, British Airways can tailor their marketing efforts to target these customers more effectively, providing them with personalized offers and incentives. This, in turn, can lead to increased revenue and a competitive advantage in the market.

Overall, this term paper seeks to address a critical business problem faced by British Airways and the wider airline industry by developing a predictive model that can accurately predict customer booking behavior. the project aims to provide practical insights and guidance that can be applied to drive growth and profitability in the highly competitive airline industry.

1.4 Project Stakeholders

1. British Airways: As the focus of the project, British Airways has a significant stake in the development of a predictive model that accurately identifies customers who are most likely to book a holiday with the airline. The successful implementation of such a model can lead to increased revenue and competitive advantage for the airline.

2. Customers: The development of a predictive model can also benefit customers by providing them with more personalized offers and incentives tailored to their travel preferences.

3. Other airlines: Other airlines in the industry may also have an interest in the development of a predictive model, as it can provide insights into the factors that influence customer booking behavior and help them optimize their own marketing strategies.

4. Regulators: Regulatory bodies may also have a stake in the project, as the use of customer data in predictive models may raise privacy and security concerns that need to be addressed.

5. Academia: Academics and researchers in the fields of machine learning, data science, and marketing may also have a stake in the project, as it can provide insights and practical applications in these areas.

2. Chapter Two: Related Work and Background Research

2.1 Literature Review

In recent years, the application of machine learning in customer buying behavior prediction has become increasingly popular. With the abundance of customer data, machine learning algorithms can analyze customer behavior and predict future buying patterns with high accuracy. In this literature review, we will examine some of the recent research on this topic.

A study by H Valecha et al. (2018)[1] aimed to predict the behavior of Consumers using a random forest algorithm and to examine the relationship between consumer behavior parameters and willingness to buy. the paper proposed a time-evolving random forest classifier that leverages unique feature engineering to predict the behavior of consumers that affects the choice of purchasing the product significantly. The results of the random forest classifier are more accurate than other machine learning algorithms.

Neha Chaudhuri et al.[2] (2021) conducted a study on the use of deep learning for predicting online customers' purchase behavior for an e-commerce platform. The study utilized a neural network with multiple hidden layers to predict customer churn. The results of the study showed that the deep learning approach outperformed traditional machine learning algorithms such as logistic regression and decision trees in predicting customer churn.

Another study by Kiran Chaudhary et al.[3] (2021) have proposed a mathematical and machine learning-based predictive model to find consumer behavior toward products on the social media platform. In this paper, they used the concept of big data technology to process data and analyze the data. The results showed that the decision tree is the best model for consumer behavior prediction on social media with an accuracy of 98% on validation data and the highest consumer deviation is 99.51% from one social media to another and the minimum is 12.22%.

In the airline industry, a study by Yoo and Lee [4] (2019) used a gradient boosting machine (GBM) to predict customer behavior. The study aimed to identify the factors that influence customers to purchase additional products such as baggage or seat upgrades. The results showed that GBM outperformed other machine learning algorithms such as logistic regression and random forests.

In summary, machine learning techniques have shown great potential for predicting customer behavior in various industries. The studies reviewed in this literature review demonstrate the effectiveness of different machine learning algorithms in predicting customer behavior. These findings provide a foundation for further research in the field and suggest the potential for using machine learning to improve customer targeting and marketing strategies.

2.2 Background and Theory

This section provides an overview of the machine learning algorithms and techniques used in this project, as well as any relevant data preprocessing or feature engineering methods. The machine learning algorithms used in this project include Logistic Regression, XGBoost, and CatBoost. These algorithms were chosen due to their proven effectiveness in classification tasks and their ability to handle large datasets.

To preprocess the data, we will use techniques such as one-hot, Catboost encoding, and robust scaling. One-hot encoding will be used to convert categorical variables into numerical data, Catboost encoding is designed for handling high-cardinality categorical features, which are those with a large number of distinct values and Robust scaling will be used to standardize the range of the features when the dataset contains outliers or

extreme values that can significantly affect the mean and standard deviation of the features.

ADASYN (Adaptive Synthetic Sampling)[5] is used for handling imbalanced datasets, where the distribution of classes is significantly skewed. Feature engineering techniques will also be employed to enhance the predictive power of the model. These techniques may include the creation of new features based on existing data, the removal of irrelevant features, and the transformation of existing features into a more suitable format for the algorithms used.

2.3 Summary

In summary, this chapter has provided a comprehensive review of the existing literature and models related to customer buying behavior prediction using machine learning in the airline industry. Additionally, it has covered the background and theory of the machine learning algorithms and techniques used in this project, as well as the relevant data preprocessing and feature engineering methods. This knowledge will serve as a foundation for the development of our predictive model in the following chapters.

3. Chapter Three: Methodology

3.1 Introduction

This chapter provides a detailed overview of the methodology used in this study to predict customer buying behavior in the airline industry using machine learning. In this section, we will provide a comprehensive overview of the data collection process, the sources of data, the data preparation steps, feature engineering, modeling, and evaluation.

The section starts with a description of the data sources, including the type of data, the size of the dataset, and the methods used to collect it. We will then outline the steps taken to prepare the data for analysis, including cleaning, feature selection, and encoding of categorical variables. We will also describe the feature engineering techniques used to derive new features from the existing ones.

The modeling approach will be presented in detail, including the selection of the algorithms and the hyperparameter tuning process. Finally, we will evaluate the performance of the models using a range of metrics and discuss the implications of the findings.

Finally, we will evaluate the performance of the models using a range of metrics and discuss the implications of the findings. Overall, this chapter provides a comprehensive account of the methodology used in this study, which is essential for understanding the results and drawing valid conclusions.

3.2 Data

The data used in this study was obtained from British Airways, one of the largest airlines in the United Kingdom. The dataset contains information on customer bookings, including flight dates, destinations, and ticket prices. The data spans several years and includes bookings made through various channels, such as the company website, travel agencies, and call centers.

Understanding the structure and properties of the data is essential for

developing an effective machine-learning model. In this chapter, we will explore the characteristics of the data and how they may impact the predictive performance of the model.

The dataset used in this study consists of 5000 rows and 14 columns. The data was collected from a travel agency and includes information on flight bookings made by customers. The following section provides a detailed description of each column in the dataset:

- num_passengers: This column indicates the number of passengers traveling for each booking.
- sales_channel: This column represents the sales channel through which the booking was made. The possible values are online, offline, and mobile.
- trip_type: This column indicates the type of trip made by the customer. The possible values are Round Trip, One Way, and Circle Trip.
- purchase_lead: This column represents the number of days between the travel date and the booking date.
- length_of_stay: This column indicates the number of days spent at the destination.
- flight_hour: This column indicates the hour of flight departure.
- flight_day: This column represents the day of the week on which the flight departed.
- route: This column represents the origin and destination flight route.
- booking_origin: This column indicates the country from where the booking was made.
- wants_extra_baggage: This column indicates whether the customer wanted extra baggage in the booking. The possible values are Yes and No.

- wants_preferred_seat: This column indicates whether the customer wanted a preferred seat in the booking. The possible values are Yes and No.
- wants_in_flight_meals: This column indicates whether the customer wanted in-flight meals in the booking. The possible values are Yes and No.
- flight_duration: This column indicates the total duration of the flight in hours.
- booking_complete: This column represents a flag indicating whether the customer completed the booking. The possible values are Yes and No.

Before performing any analysis, we conducted several preprocessing steps to clean and prepare the data for modeling. These steps included removing any missing values, encoding categorical variables, and scaling the features. The details of the data preparation and feature engineering steps are discussed in Section 3.4.

3.3 Exploratory Data Analysis

In this section, we will explore the dataset to gain a better understanding of the variables and their relationships. We will perform a variety of analyses, including summary statistics, visualizations, and correlation analysis.

3.3.1 Summary Statistics

We will start by calculating summary statistics for each variable in the dataset. This will give us a general idea of the distribution of the data and any outliers that may be present. We will calculate basic statistics such as mean, median, mode, standard deviation, and range. These statistics can provide

insights into the data distribution, central tendency, and dispersion.

index	count	mean	std	min	25%	50%	75%	max
wants_preferred_seat	50000.0	0.29696	0.45692333490278164	0.0	0.0	0.0	1.0	1.0
wants_in_flight_meals	50000.0	0.42714	0.49466788285301727	0.0	0.0	0.0	1.0	1.0
wants_extra_baggage	50000.0	0.66878	0.47065671349173294	0.0	0.0	1.0	1.0	1.0
purchase_lead	50000.0	84.94048	90.45137813436415	0.0	21.0	51.0	115.0	867.0
num_passengers	50000.0	1.59124	1.020164730385021	1.0	1.0	1.0	2.0	9.0
length_of_stay	50000.0	23.04456	33.887670056969654	0.0	5.0	17.0	28.0	778.0
flight_hour	50000.0	9.06634	5.412659692064414	0.0	5.0	9.0	13.0	23.0
flight_duration	50000.0	7.2775608	1.496862916327065	4.67	5.62	7.57	8.83	9.5
booking_complete	50000.0	0.14956	0.35664316941027446	0.0	0.0	0.0	0.0	1.0

Table 1 Descriptive statistics

Table 1 shows the descriptive statistics of the numerical features in the dataset. It can be observed that Most of the numerical columns have a right-skewed distribution and The data has variable scales so we will normalize it later.

index	count	unique	top	freq
sales_channel	50000	2	Internet	44382
trip_type	50000	3	RoundTrip	49497
flight_day	50000	7	Mon	8102
route	50000	799	AKLKUL	2680
booking_origin	50000	104	Australia	17872

Table 2 Frequency distribution

Table 2 shows the frequency distribution of the categorical features in the dataset. It can be observed that the most common sales channel used for

booking was online, and the most common trip type was a round trip. Additionally, it can be observed that most bookings were made from Australia.

```
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   num_passengers         50000 non-null  int64
 1   sales_channel          50000 non-null  object
 2   trip_type              50000 non-null  object
 3   purchase_lead          50000 non-null  int64
 4   length_of_stay         50000 non-null  int64
 5   flight_hour            50000 non-null  int64
 6   flight_day             50000 non-null  object
 7   route                  50000 non-null  object
 8   booking_origin         50000 non-null  object
 9   wants_extra_baggage    50000 non-null  int64
 10  wants_preferred_seat   50000 non-null  int64
 11  wants_in_flight_meals  50000 non-null  int64
 12  flight_duration        50000 non-null  float64
 13  booking_complete       50000 non-null  int64
dtypes: float64(1), int64(8), object(5)
memory usage: 5.3+ MB
```

Info function

The info() function provides us with information on the number of non-null values and data types of each feature in our dataset. From the output, we can see that there are 5000 entries in the dataset and 14 features. None of the features have any missing values, as indicated by the non-null count for each feature. The data types for the features are a mix of integer, float, and object, which suggests that there are both numerical and categorical features in the dataset. Some of the columns should be converted into different data types, e.g. flight_day. Finally, it seems that the dataset contains 719 duplicate values.

3.3.2 Data Visualization

In addition to summary statistics, data visualization is an important tool

for understanding the distribution of the data and identifying any patterns or relationships between variables. In this section, we will present some visualizations of the dataset.

3.3.2.1 Visualize target column

To begin, let's examine the distribution of the target column, which represents whether a customer completed the booking. Figure 1 shows a count plot of the booking_complete feature.

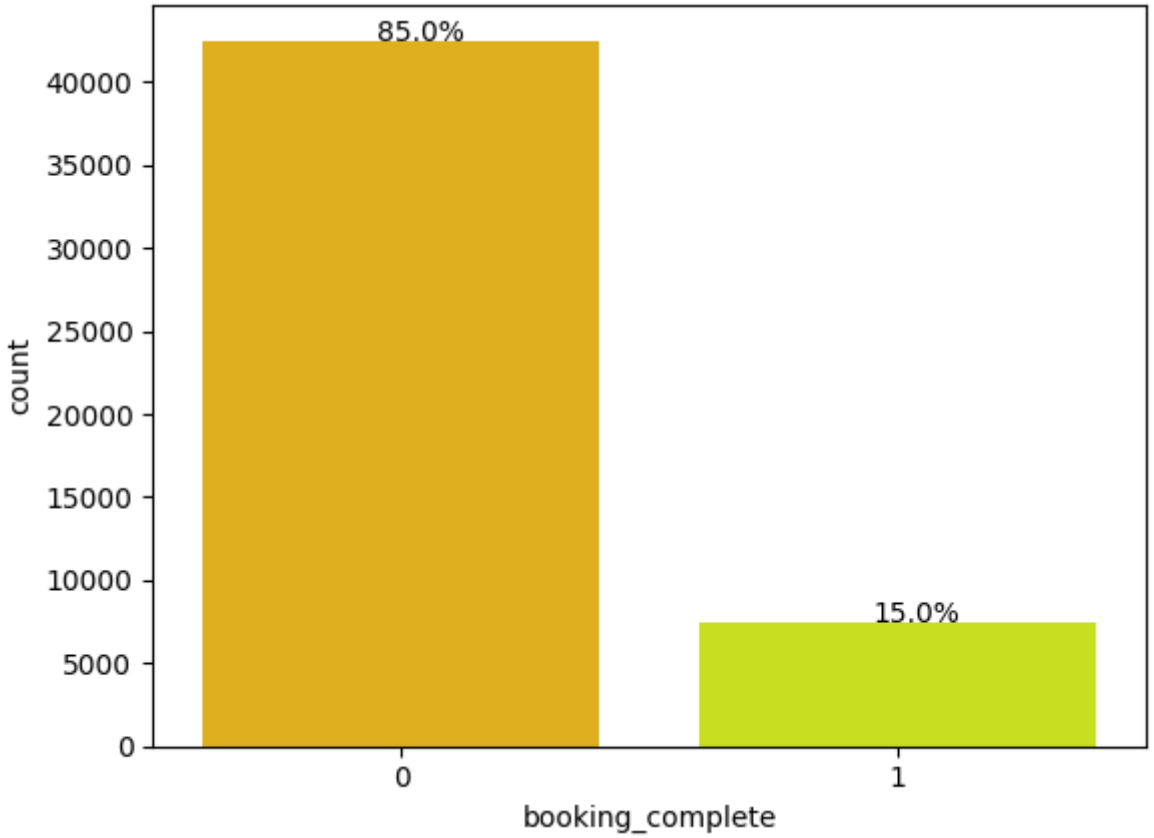

Figure 1 Count plot of Target column

The count plot provides an overview of the distribution of the target variable. From the plot, we can observe the following:

1. About 15% of the total customers have completed the booking.
2. The dataset is "imbalanced" because the distribution of the target

variable is not equal across different classes.

3.3.2.2 Visualize Categorical Features

Next, let's explore the categorical features in the dataset. We can utilize various visualizations, such as bar plots, count plots, and pie charts, to gain insights into the distribution and frequencies of these features. Here are a few examples:

3.3.2.1.1 Sales Channel

Figure 2 presents a count plot depicting the distribution of bookings across different sales channels.

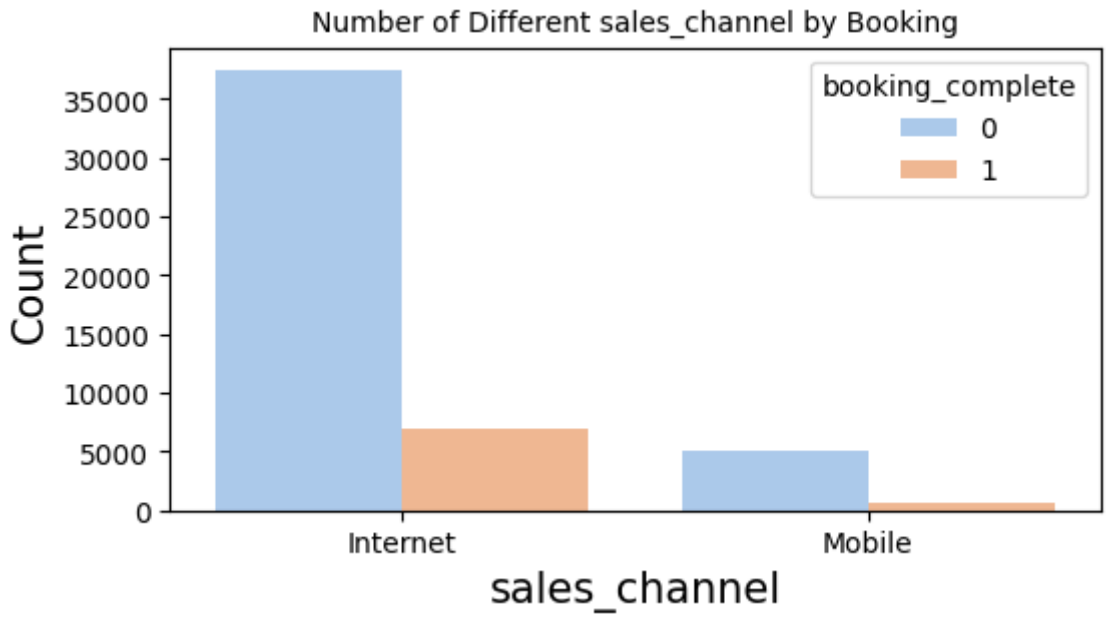

Figure 2 Count plot of sales_channel

From the plot, we can observe the following:

- The number of consumers who finished their booking via the Internet is bigger than the number of customers who completed their booking via mobile.

3.3.2.1.2 Trip Type

Figure 3 showcases a count plot illustrating the proportion of different trip types in the dataset.

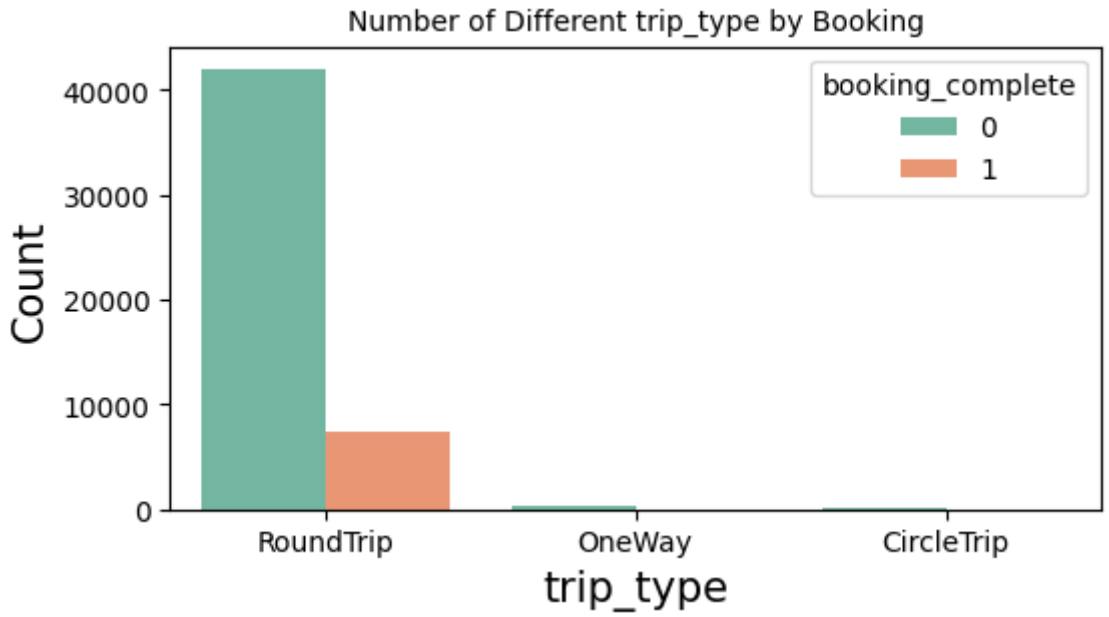

Figure 3 Count plot of trip_type

From the plot, we can observe the following:

- The number of customers buying Round trips is significantly higher than the number of people booking other trip types.
- In the Round trip type, the number of customers who have not completed their booking is nearly three times the number of customers who have completed their booking.

3.3.2.1.2 Flight day

Figure 4 showcases a count plot illustrating the distribution of bookings across different days of the week.

The count plot helps us understand the distribution of bookings across different days of the week. It provides insights into any variations in booking patterns based on weekdays or weekends.

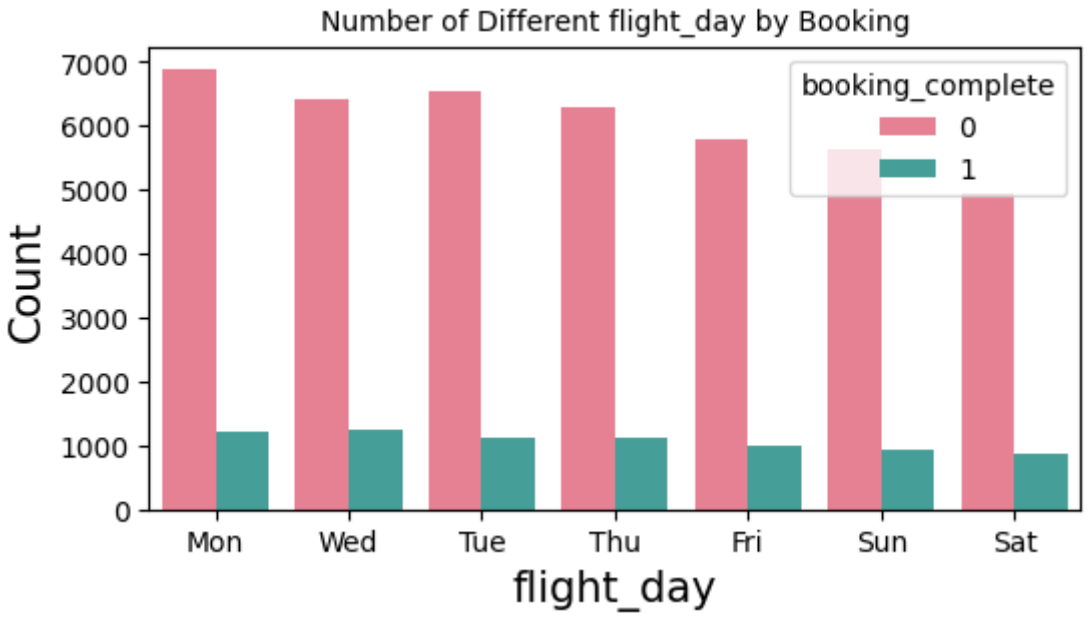

Figure 4 Count plot of flight_day

From the previous plot, It looks like that on weekends the number of passengers that have completed their booking is less than on weekdays. So let's check that by looking at the average number of passengers per day.

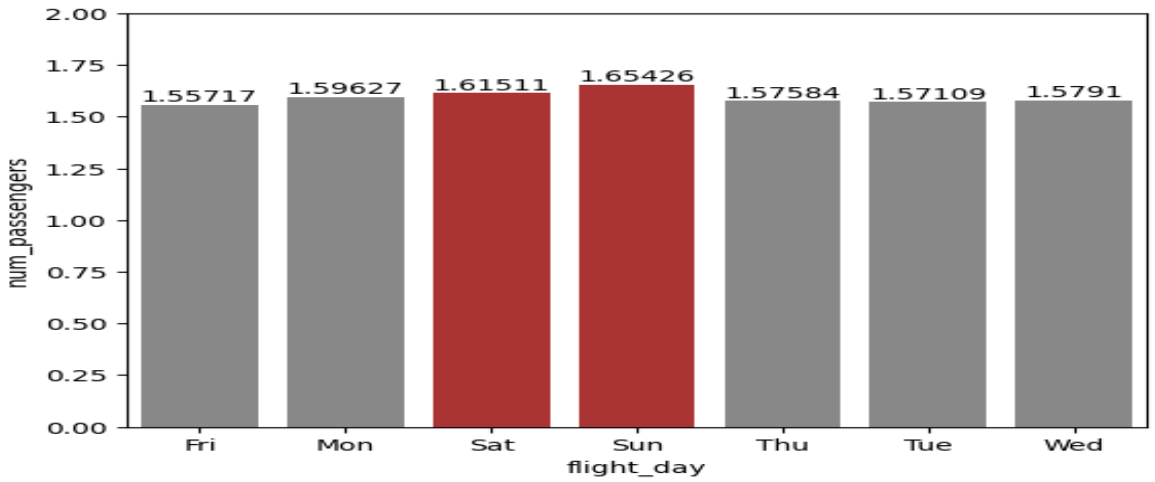

Figure 5 flight_day VS. average number of passengers

On weekends the number of passengers is less than on weekdays, but the average number of passengers, on weekends has a higher value than on weekdays, so we need to see the number of flights from day to day.

	index	flight_day
4	Fri	6761
0	Mon	8102
6	Sat	5812
5	Sun	6554
3	Thu	7424
2	Tue	7673
1	Wed	7674

Table 3 the number of flights

From Table 3, we can see that weekends have fewer flights than weekdays, so the company needs to consider adding flight schedules on weekends.

3.3.2.1.3 Route

Figure 5 displays the routes that have the highest number of passengers

on weekends.

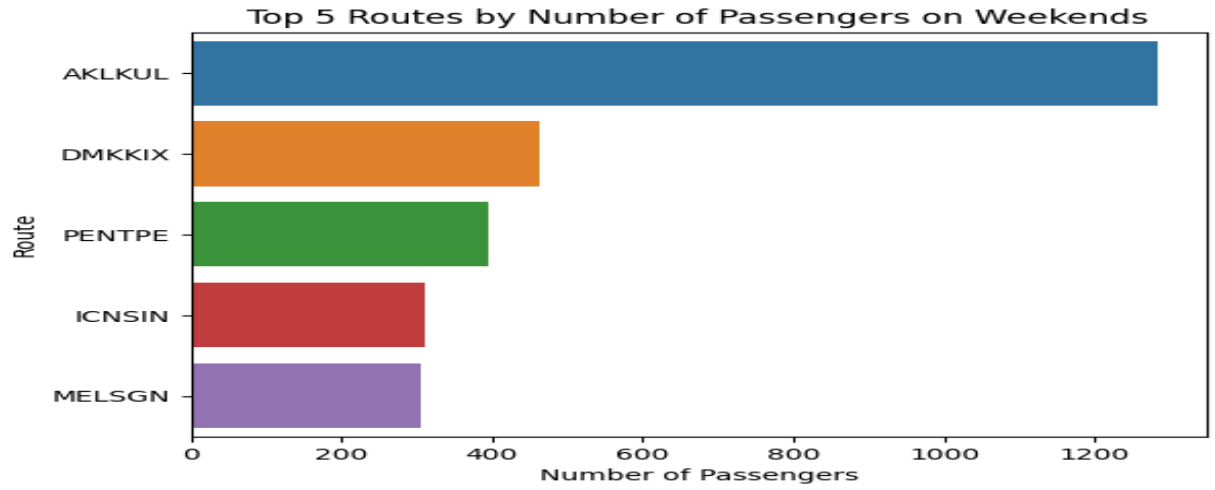

Figure 6: Bar plot of Top 5 Routes

From the plot above we can see the top 5 routes that have the most passengers, so it's recommended that the company increases the number of flights to these five routes on weekends.

3.3.2.1.4 Booking origin

Figure 6 displays the Top Ten Countries by Bookings.

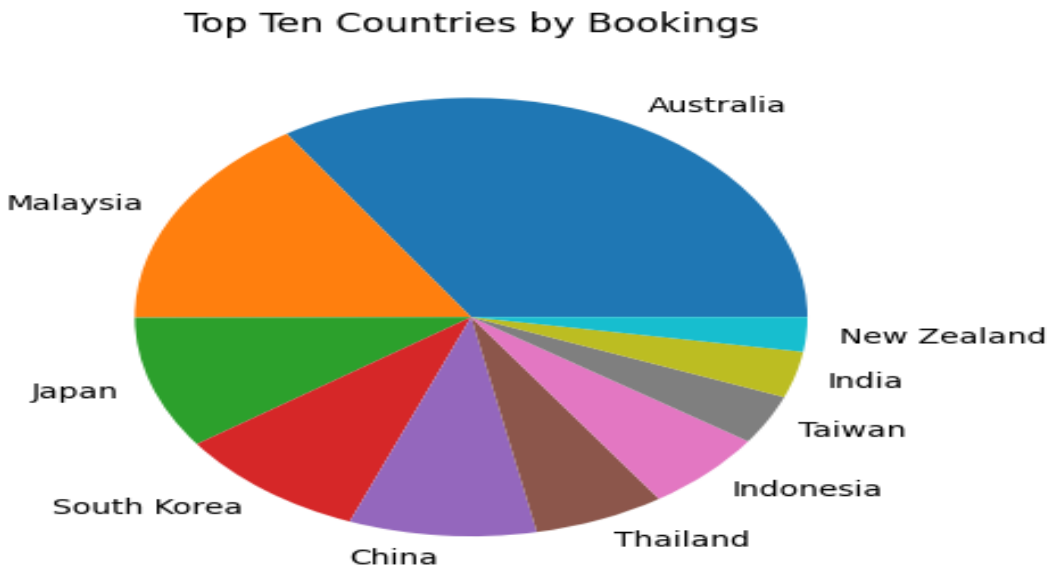

Figure 7 Bie plot of Top 10 Countries by Bookings

Apparently, Australia has the most bookings in this dataset.

3.3.2.3 Visualize Binary Features

Additionally, we can explore the distribution and patterns of binary features.

3.3.2.3.1 Wants extra baggage, preferred seats, flight meals

Figure 7 showcases multiple bar plots illustrating the proportion of customers who wanted extra baggage, preferred seats, and flight meals in their bookings.

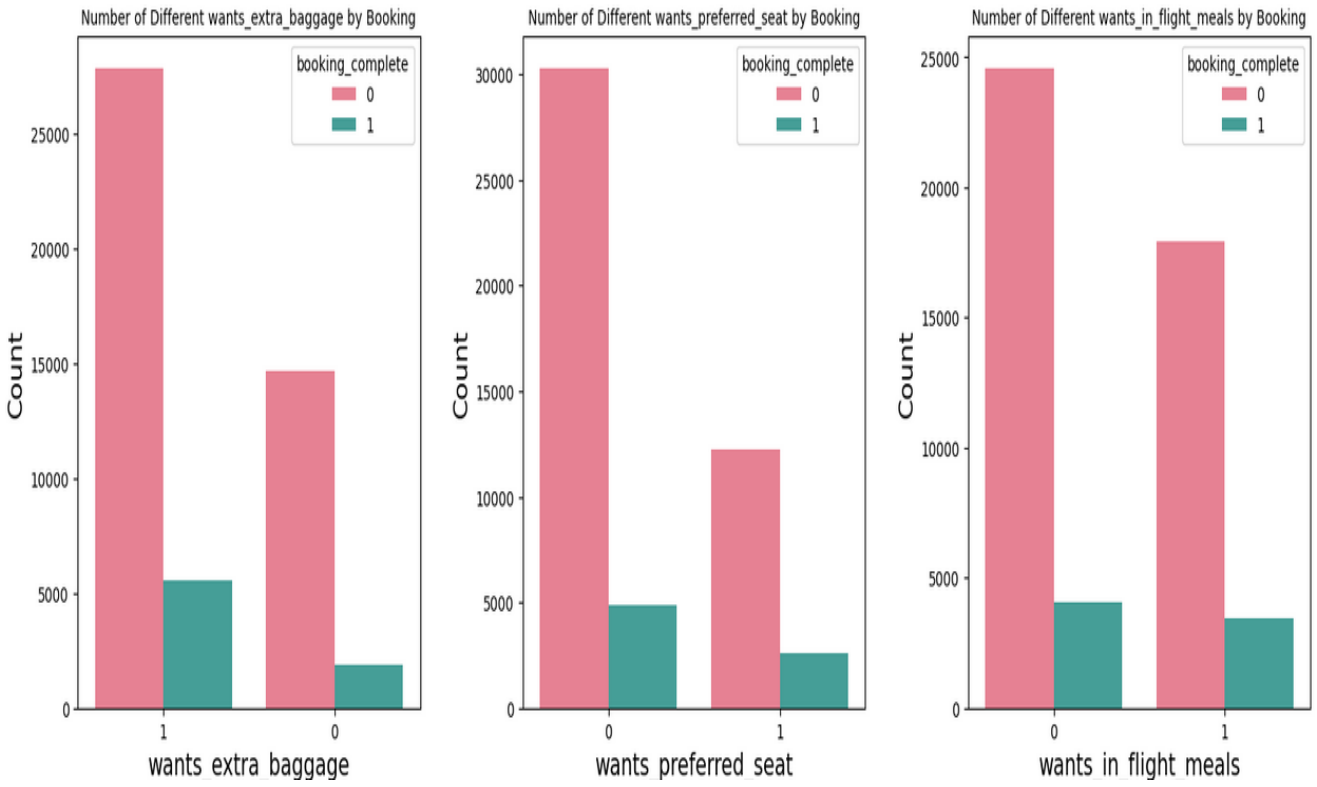

Figure 8 Bar plot of binary features

From the plot, we can observe the following:

- Passengers who want extra baggage and a premium seat have nearly double the completed booking rate as those who do not.

- The wants in-flight meal column doesn't have significant predictive power. A similar percentage of booking is shown both when a customer wants a meal or not.

3.3.2.4 Visualize Numerical Features

In addition to categorical features, it is crucial to explore the distribution and relationships of numerical features. Visualizations such as KDE plots, box plots, and scatter plots can be employed. Let's consider a few examples:

3.3.2.4.1 KDE plots

To gain a deeper understanding of the numerical features, we can utilize Kernel Density Estimation (KDE) plots. Let's consider the following numerical features: 'num_passengers', 'purchase_lead', 'length_of_stay', 'flight_hour', and 'flight_duration'. Figures 8-11 showcase the KDE plots for these features.

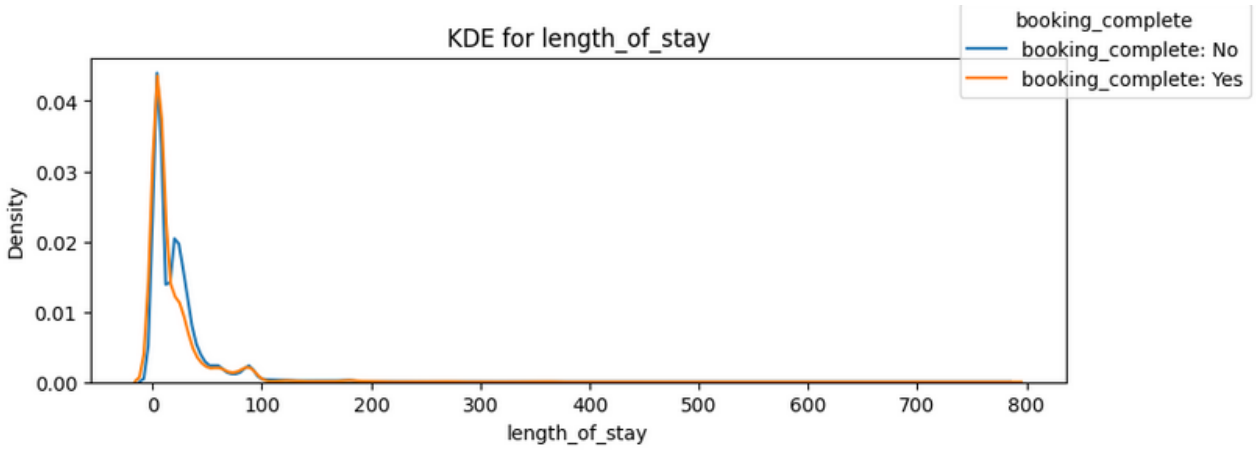

Figure 9 KDE plot of length_of_stay

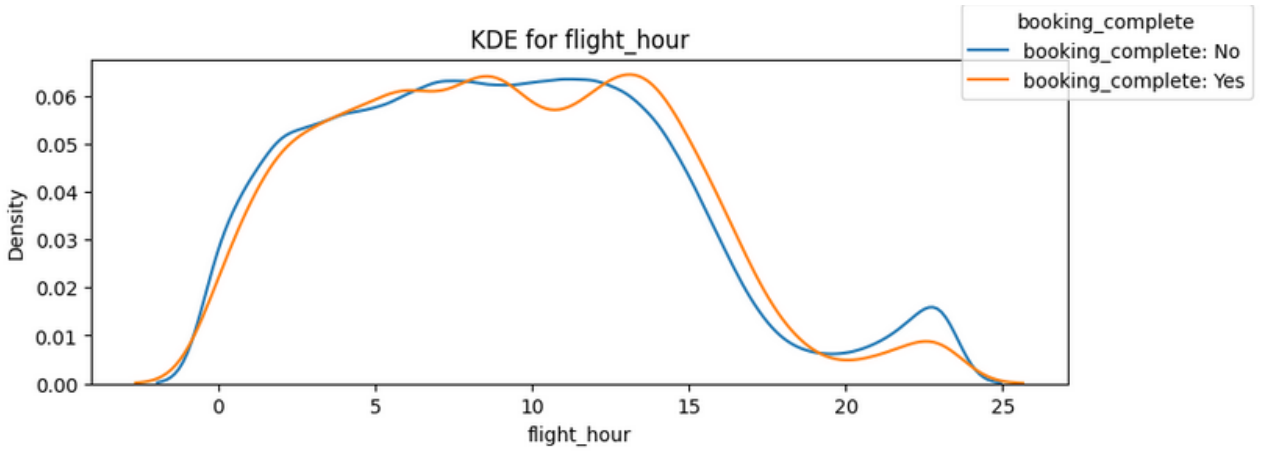

Figure 10 KDE plot of flight_hour

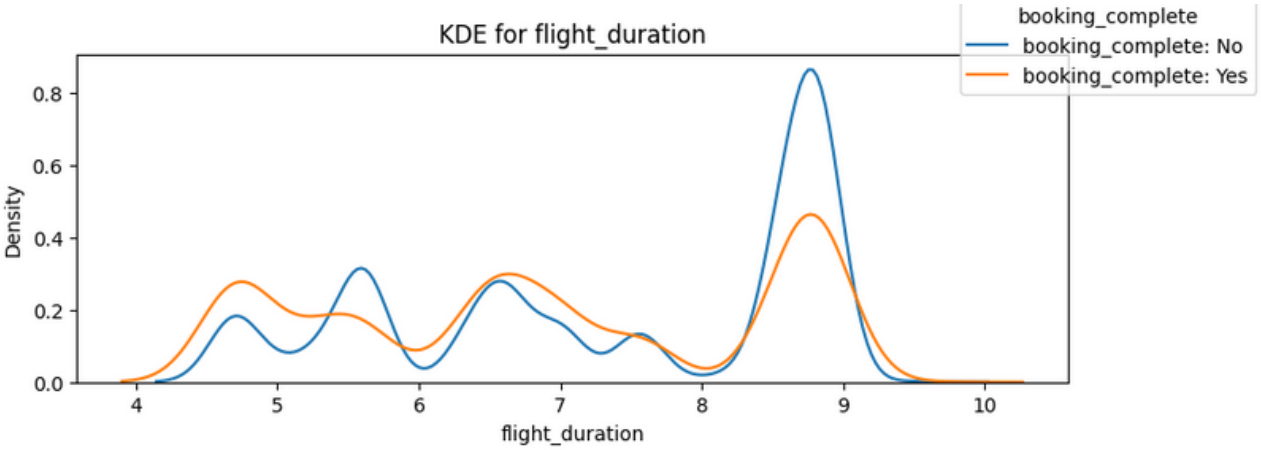

Figure 11 KDE plot of flight_duration

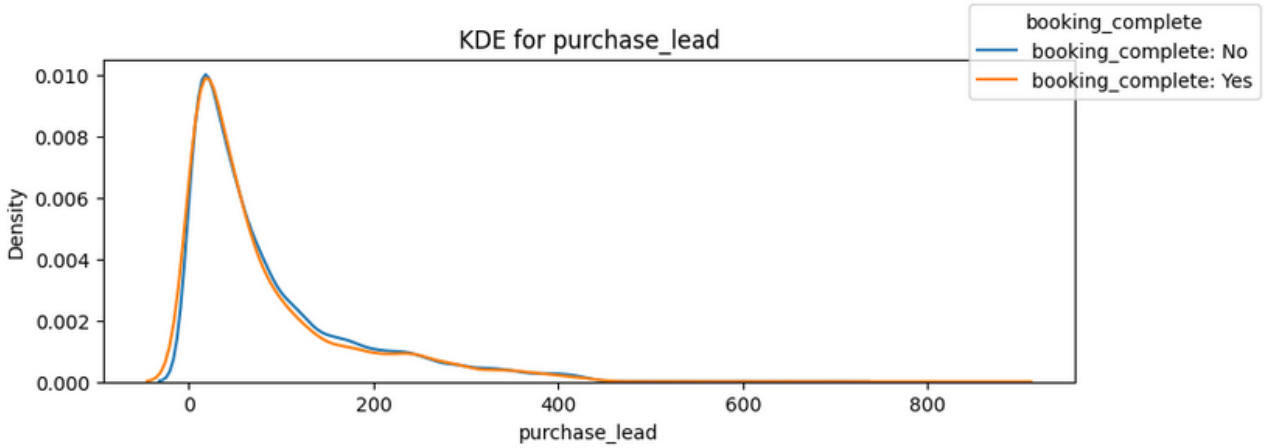

Figure 12 KDE plot of purchase_lead

The KDE plots provide insights into the distribution and density of values for each numerical feature. They can help us identify the presence of multiple peaks, outliers, and potential relationships between the features.

From the above plots, We can conclude the following:

- The greater the number of days between the date of travel and the date of booking(purchase_lead), the lower the percentage of completion of the booking.
- The higher the number of days spent at the destination(length_of_stay), the lower the proportion of the booking completed.
- The higher the total duration of the flight (in hours)(flight_hour), the lower the proportion of the booking completed.
- The later the flight departure hour(flight_duration), the lower the proportion of the booking completed.

3.3.2.4.2 Box plots

To gain a deeper understanding of the outliers in numerical features, we can utilize box plots. Let's consider the following numerical features: 'num_passengers', 'purchase_lead', 'length_of_stay', 'flight_hour', and 'flight_duration'. Figure 12 showcases the box plots for these features.

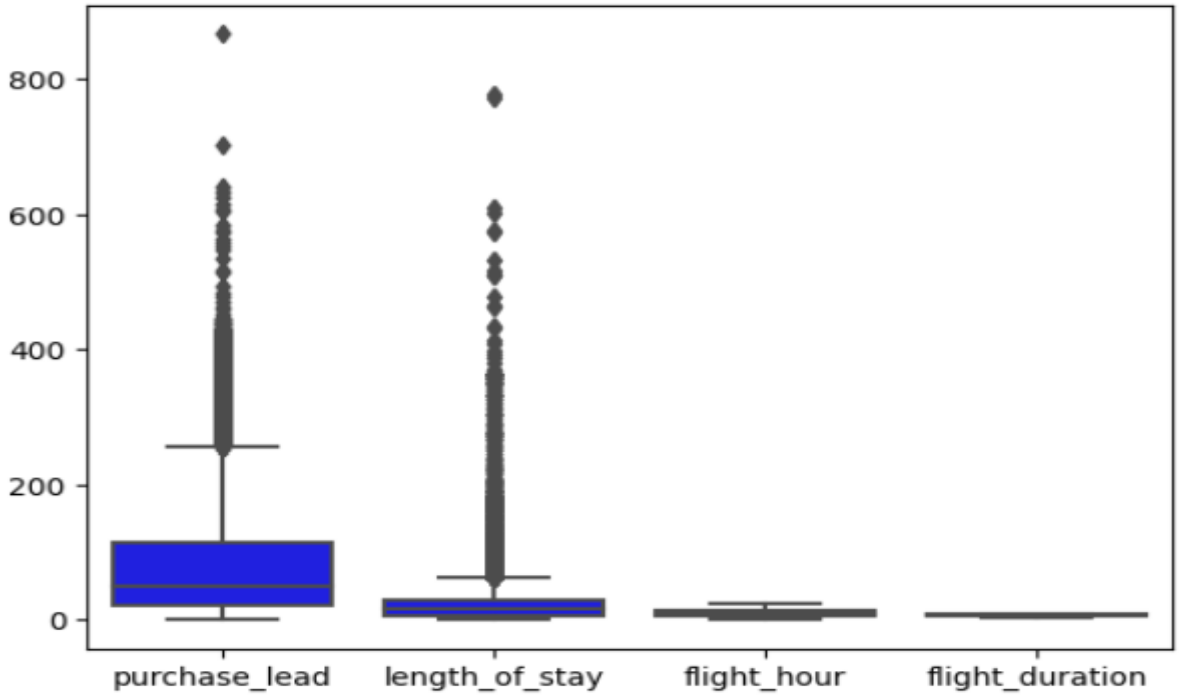

Figure 13 Box plot of numerical features

From Figure 12, We can conclude the following:

- The numerical columns require normalization.
- The purchase lead and the length of stay columns have outliers.

By utilizing a combination of visualizations, we can gain a comprehensive understanding of the dataset and uncover valuable insights. These visualizations will serve as a foundation for further analysis and modeling in subsequent chapters.

3.3.3 Correlation Analysis

Correlation analysis is a valuable technique used to measure the strength and direction of relationships between variables in the dataset. By examining the correlations, we can gain insights into the interdependencies among different features and identify potential patterns or associations. In this section, we will explore the correlation between features in our dataset.

3.3.3.1 Pearson Correlation

The Pearson correlation[6] coefficient measures the linear relationship between two continuous variables. It ranges from -1 to 1, where a value close to 1 indicates a strong positive correlation, a value close to -1 indicates a strong negative correlation and a value close to 0 indicates no significant correlation.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

To investigate the Pearson correlation between numerical and binary features, we calculated the correlation matrix for 'num_passengers', 'purchase_lead','length_of_stay','flight_hour','wants_extra_baggage','wants_preferred_seat',' wants_in_flight_meals',' booking_complete' and 'flight_duration'. Figure 13 showcases the correlation matrix.

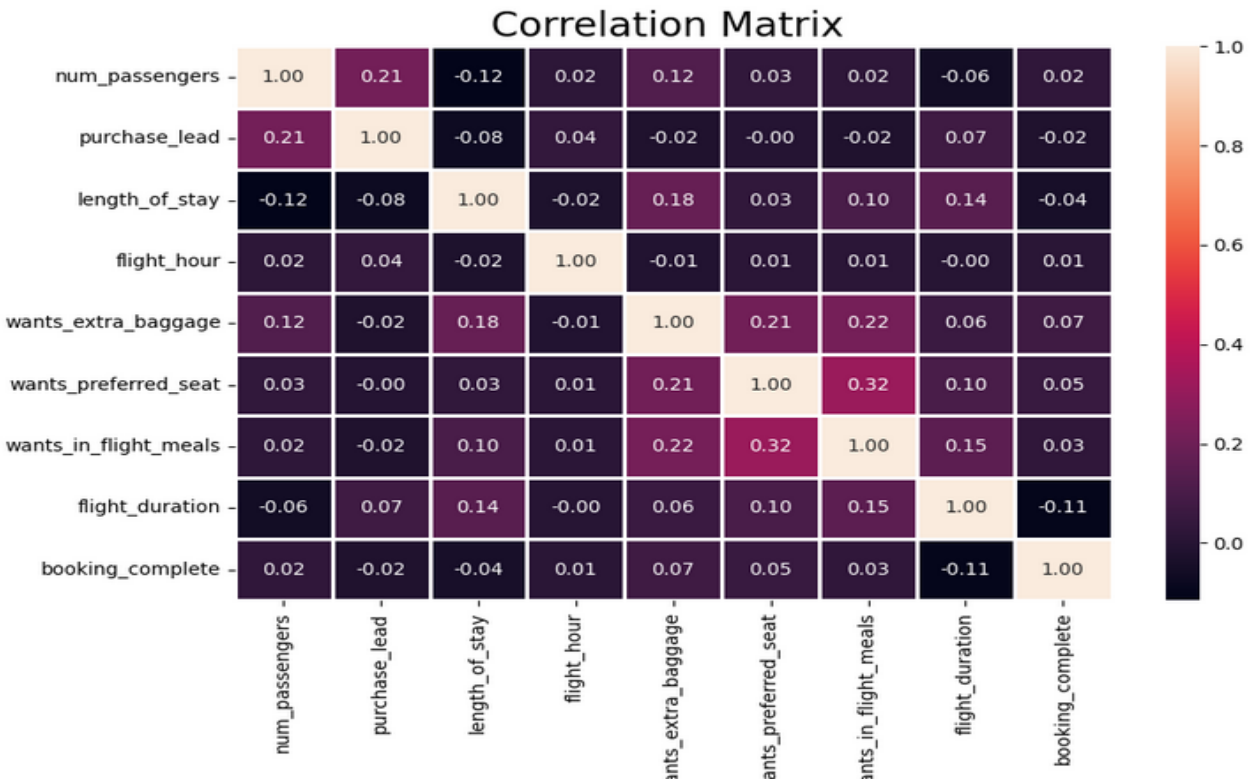

Figure 14 Correlation matrix

From the correlation matrix, we can observe that there is no multicollinearity among the numerical independent variables. This suggests that these numerical features are relatively independent and do not exhibit high correlations that could affect the stability and interpretability of our models.

3.3.3.2 Mutual Information

Mutual information[7] is a measure of the mutual dependence between two variables, regardless of whether the relationship is linear or not. It quantifies the amount of information that one variable provides about the other. In the context of feature selection, mutual information can help identify the relevance of a feature in predicting the target variable. Higher values of mutual information show a higher degree of dependency which indicates that the independent variable will be useful for predicting the target.

Features	Importance
route	0.060984
booking_origin	0.045508
sales_channel	0.000908
trip_type	0.000516
flight_day	0.000159

Table 4 Feature Importance

Based on the previous values in Table 4, it seems that the "route" and "booking_origin" features have the highest mutual information with the target variable, indicating a stronger relationship. The "sales_channel" and "trip_type" features have lower mutual information values, indicating a

weaker relationship. The "flight_day" feature has the lowest mutual information value, indicating almost no relationship with the target variable.

3.4 Data Preprocessing

Data preprocessing is a crucial step in preparing the dataset for analysis and modeling. This section will address various preprocessing tasks, including handling missing values, removing duplicates, and converting data types.

3.4.1 Handling Missing Values

Upon examining the dataset, we found that there are no missing values in any of the columns. This is a positive indication as it allows us to proceed with the analysis without the need for imputation or interpolation techniques.

3.4.2 Removing Duplicates

During our data exploration, we identified the presence of 719 duplicate values in the dataset. These are exact duplicates, where two or more records have identical values for all variables. To ensure the integrity and accuracy of our analysis, we will remove these duplicate entries from the data frame. We will retain the first occurrence of each row and discard the subsequent duplicates.

3.4.3 Converting Data Types

In some cases, it may be necessary to convert specific columns to different data types to enhance their utility for analysis and modeling. For example, the 'flight_day' column represents the day of the week for flight departures. To facilitate further analysis, we can convert these weekdays into

numeric values.

To achieve this, we will utilize a mapping dictionary that assigns a numeric value to each weekday. This conversion will enable us to leverage the ordinal nature of the weekdays in subsequent calculations or modeling tasks.

Finally, By addressing these data preprocessing steps, we ensure the quality and consistency of the dataset, setting a solid foundation for subsequent analyses and predictive modeling.

3.5 Feature Engineering

Feature engineering is an essential step in preparing the dataset for modeling. In this section, we will discuss the feature engineering techniques applied to enhance the predictive power of the dataset.

3.5.1 Feature Engineering

During the feature engineering process, we created a new column called 'is_weekend' to capture whether the flight day falls on a weekend. This new column will provide valuable information about the relationship between the flight day and customer behavior.

3.5.2 Encoding

We applied one-hot encoding[8] to convert categorical variables with more than two levels such as 'sales_channel' and 'trip_type' into binary indicator variables. This transformation allows the inclusion of categorical information in our models effectively.

However, the 'route' and 'booking_origin' columns have multiple

distinct values, which would lead to a high-dimensional feature space if we applied one-hot encoding with them. To address this issue, we used the CatBoost encoder[11] for these columns. The CatBoost encoder effectively encodes categorical variables with many levels while maintaining model performance.

3.5.3 Oversampling the Minority Class using ADASYN

The dataset exhibits an imbalanced distribution, where the number of instances in the minority class is significantly lower than in the majority class. Training a machine learning model on such imbalanced data can lead to poor performance and biased predictions. In situations where collecting more data is not feasible, we must employ techniques to address this class imbalance issue.

When dealing with imbalanced datasets, traditional approaches like under-sampling, which removes samples from the majority class, can result in the loss of valuable information that could have been used to train the model effectively. On the other hand, random oversampling blindly duplicates instances from the minority class, which can cause the model to overfit and perform poorly on unseen data.

To overcome these limitations, we employed the ADASYN (Adaptive Synthetic) technique, which is specifically designed to address class imbalance. ADASYN increases the sample size of the minority class by generating synthetic samples based on the existing minority instances. Unlike random oversampling, ADASYN adapts the synthetic sample generation process based on the density distribution of the minority class, focusing more on challenging regions where the class overlap is higher.

By incorporating ADASYN, we effectively increase the representation

of the minority class in the dataset without overfitting the model. This technique helps to mitigate the impact of class imbalance, enabling the model to learn from a more balanced set of examples and improve its predictive performance on both the minority and majority classes. After We applied ADASYN to our dataset we compared the results before and after using smote. The results before and after ADASYN:

Class	Precision	Recall	F1-score
0	0.86	0.97	0.91
1	0.45	0.12	0.19

Table 5 Results before ADASYN

Class	Precision	Recall	F1-score
0	0.76	0.75	0.76
1	0.76	0.77	0.77

Table 6 Results after ADASYN

Based on the comparisons above, we can conclude that after using ADASYN, the precision value increased from 45% to 0.76%, which is a significant improvement.

3.5.4 Transformation: Log Transformation

To reduce the impact of outliers in numerical columns(purchase_lead', length_of_stay'), we applied a log transformation. This transformation helps to normalize the data and create a more symmetric distribution, which can improve the performance of certain models that assume normality.

3.5.5 Feature Scaling: RobustScaler

To ensure that features are on a similar scale and prevent the influence

of outliers, we employed the RobustScaler[9]. The RobustScaler scales the features by subtracting the median from each feature value and then dividing by the interquartile range (IQR). This process removes the median and scales the data based on the IQR, making it more robust to outliers compared to other scaling techniques.

The RobustScaler is particularly useful in scenarios where the data contains outliers or is not normally distributed. It helps to ensure that the features are on a similar scale while preserving the relative relationships between values. This can be important for machine learning algorithms that rely on the assumption of similar feature scales for accurate model training and predictions.

After applying the log transformation and RobustScaler to the numerical features, we observed a significant change in the distribution and handling of outliers. Figure 14 displays the boxplots of the transformed features.

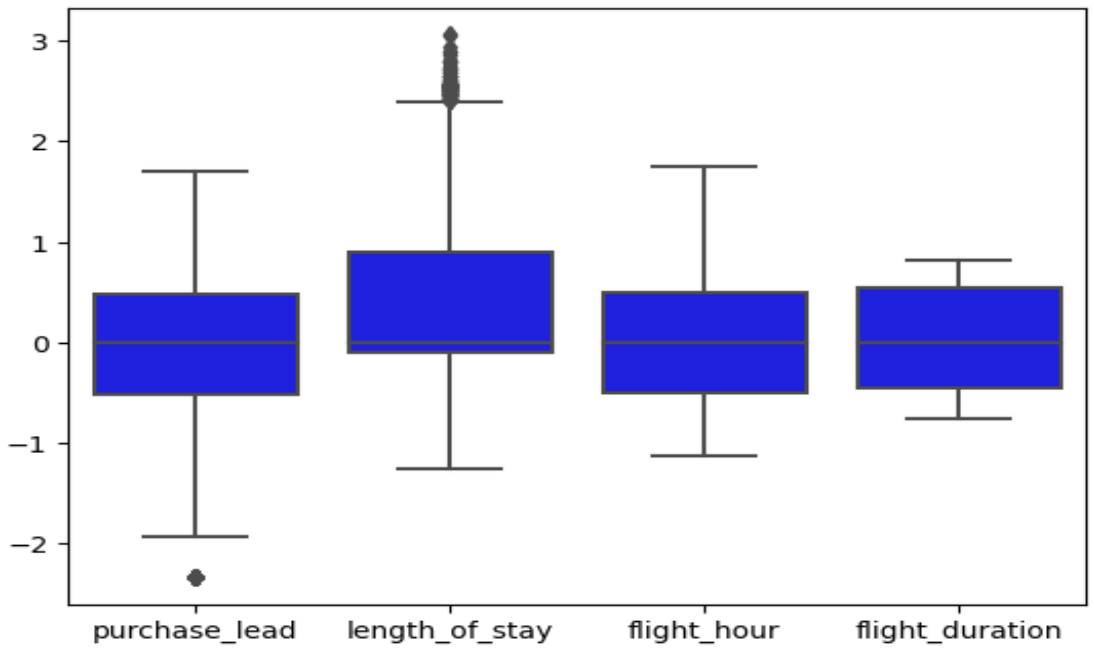

Figure 15 The boxplots of the transformed features.

In summary, the combination of log transformation and RobustScaler has effectively addressed the issues of skewed distributions and outliers in the numerical features. The resulting boxplots demonstrate a more normalized range and reduced impact of outliers, setting a solid foundation for subsequent modeling and analysis.

3.5.6 Feature Importance: Random Forest

Lastly, we employed a Random Forest classifier to determine the importance of features in predicting customer buying behavior. The Random Forest algorithm ranks the features based on their contribution to the overall model performance. This information aids in identifying the most influential features and selecting the most relevant ones for our predictive models.

Each tree of the random forest can calculate the importance of a feature according to its ability to increase the pureness of the leaves. The higher the increment in leaf purity, the higher the importance of the feature.

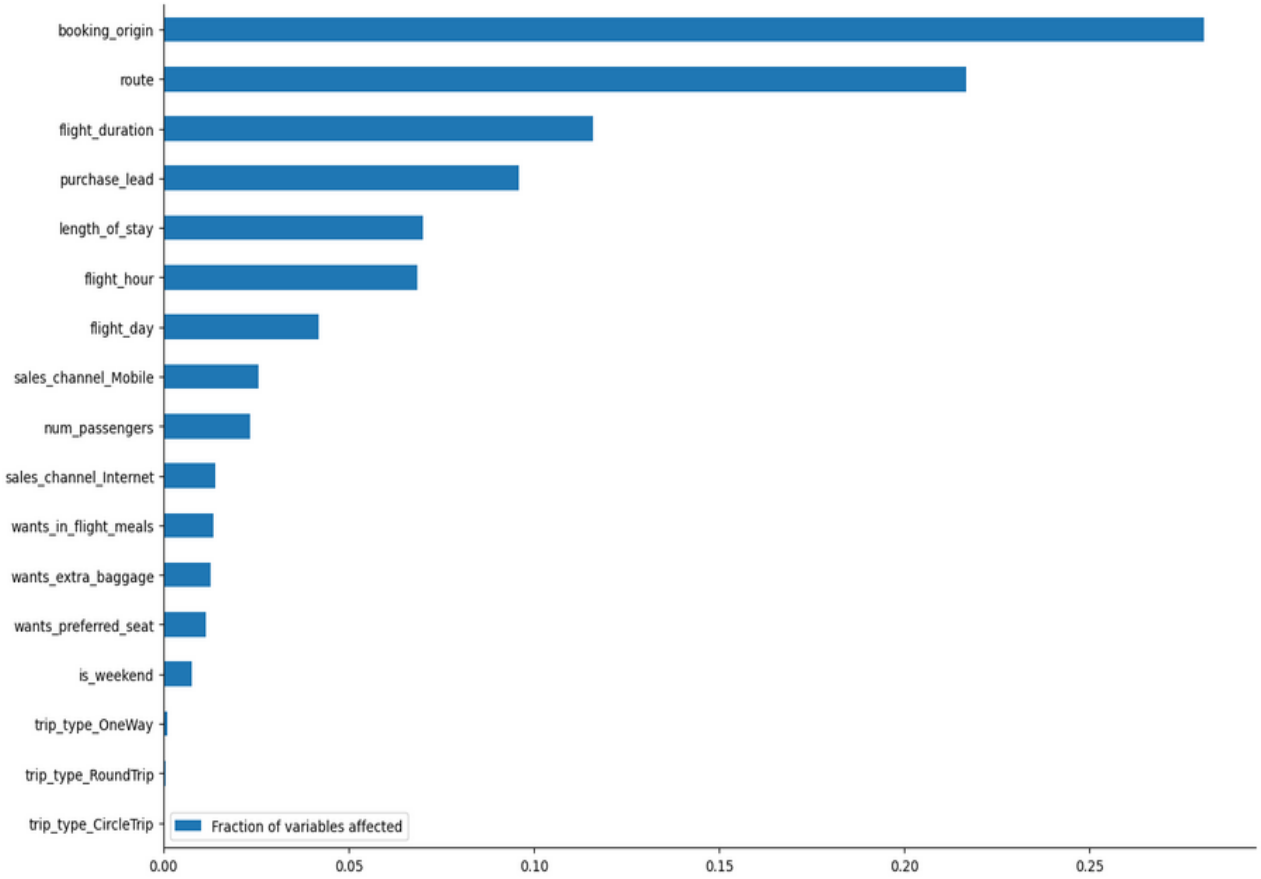

Figure 16 Feature importance with random forest.

The above graph shows that 'booking_origin' and 'route are the most important features.

3.6 Summary

This chapter has provided an in-depth exploration of the various techniques and steps involved in analyzing and preparing the data for predicting customer buying behavior. Throughout this chapter, we discussed the data collection process, data visualization, data preprocessing techniques, feature selection and engineering, and the use of machine learning algorithms for modeling. By implementing these feature selection and engineering techniques, we enhance the dataset's predictive power and improve the accuracy and interpretability of our models.

In conclusion, this chapter provides a comprehensive overview of the methodology employed to preprocess the data and engineer relevant features for predicting customer buying behavior. By addressing data quality issues, handling imbalanced data, and applying appropriate transformations and scaling techniques, we have laid a solid foundation for building robust machine learning models in the subsequent chapters.

4. Chapter Four: Modeling

4.1 Introduction

In this chapter, we present the modeling phase of our study, where we applied three different machine learning models to predict customer buying behavior. The models used include Logistic Regression (Baseline Model), XGBoost Classifier, and Catboost Classifier. We will discuss the results obtained from each model and evaluate their performance using various metrics.

4.2 Model Training

In this section, we discuss the selection of models for our study based on the problem at hand and the nature of the dataset. We considered several factors, including the classification task and the potential to capture complex patterns in the data. After careful consideration, we selected three models for our study: Logistic Regression, XGBoost Classifier, and Catboost Classifier.

4.2.1 Logistic Regression

Logistic Regression is a popular and widely used algorithm for binary classification tasks. It is known for its simplicity, interpretability, and ability to handle linear relationships between features and the target variable. We chose Logistic Regression as a baseline model to establish a performance benchmark for more complex models. The results obtained from the Logistic Regression model are as follows:

Class	Precision	Recall	F1-score
0	0.76	0.75	0.76
1	0.76	0.77	0.77

Table 7 Logistic regression results

4.2.2 XGBoost Classifier

XGBoost (Extreme Gradient Boosting) is a powerful ensemble

learning algorithm that utilizes gradient boosting techniques. It excels at capturing non-linear relationships, handling complex feature interactions, and achieving high predictive performance. We selected XGBoost Classifier for its ability to handle complex patterns in the data and deliver accurate predictions.

Class	Precision	Recall	F1-score
0	0.87	0.92	0.89
1	0.92	0.86	0.89

Table 8 XGboost results

4.2.2 Catboost Classifier

Catboost is another gradient-boosting algorithm that is specifically designed to handle categorical features efficiently. It incorporates novel strategies to deal with categorical data, such as applying numerical transformations to categorical variables and optimizing the learning process. We included Catboost Classifier in our model selection to leverage its capability in handling categorical variables effectively.

By selecting these three models, we aimed to cover a range of techniques and approaches that are well-suited for the given classification task. These models were chosen based on their strengths in capturing different types of relationships in the data and their potential to deliver accurate and reliable predictions.

Class	Precision	Recall	F1-score
0	0.87	0.93	0.91
1	0.93	0.86	0.89

Table 9 Catboost results

4.3 Model Evaluation and Selection

In this section, we assess the performance of the selected models using various evaluation metrics, including ROC curves, AUC scores, learning curves, and Confusion matrix. After training the models on the prepared

dataset, we evaluated their performance using AUC scores. The AUC scores obtained for each model are as follows:

- Logistic Regression: 0.8377747881629358.
- XGBoost: 0.9615287105758057.
- Catboost: 0.9604350597280348

The AUC score is a widely used metric for binary classification tasks, indicating the model's ability to distinguish between positive and negative instances. A higher AUC score indicates better discrimination power and overall performance.

The ROC curves depict the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity) at different classification thresholds. The closer the curve is to the top-left corner, the better the model's performance. Let's examine the ROC curve in all models:

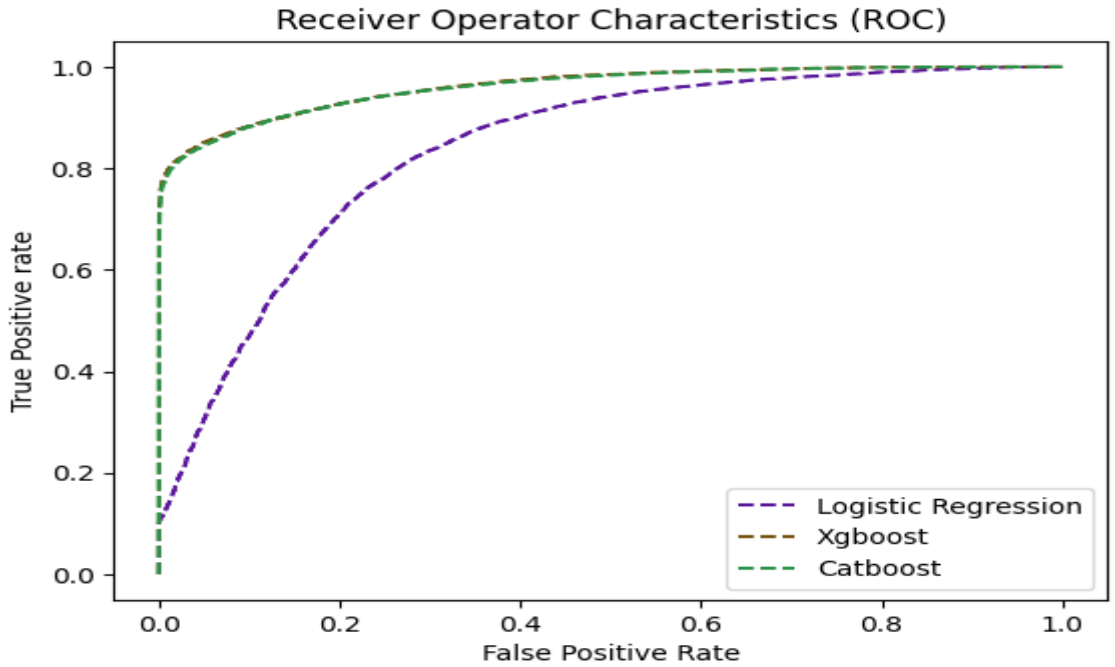

Figure 17 Roc curve for all models.

In addition to the ROC curves, we plotted learning curves to analyze the model's learning performance over time. Learning curves demonstrate changes in learning performance in terms of experience and help diagnose whether the model is underfitting, overfit, or well-fit. The figure below shows the learning curves for all applied machine-learning models:

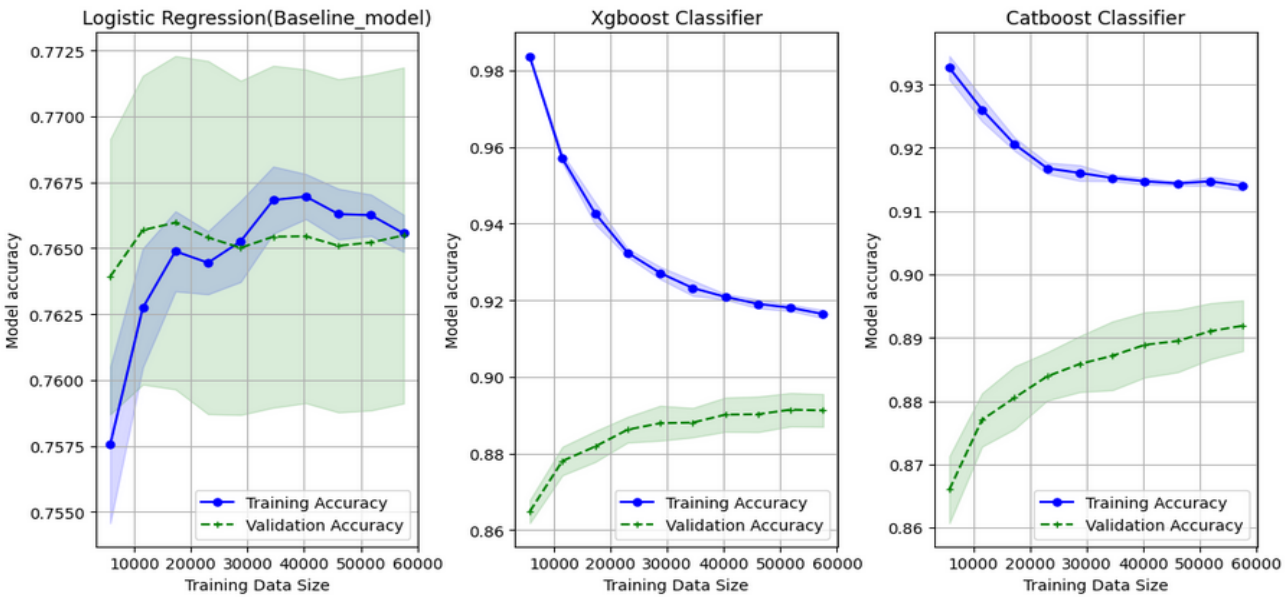

Figure 18 Learning Curves for all models.

By analyzing the learning curves, we can gain insights into how each model's performance evolves as more training examples are added. This helps us diagnose whether the model is underfitting (high bias) or overfitting (high variance).

In the evaluation of our models, we also considered the confusion matrix, which provides a detailed breakdown of the model's predictions. The confusion matrix helps us assess the performance of the model in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Dependent on the business needs and the specific context of our prediction task, the most important metric is Precision because the cost of a false positive (predicting a customer will book a holiday with British Airways when they will not) may be higher than the cost of a false negative (predicting a customer will not book a holiday with British Airways when they will). Therefore, the focus is on maximizing precision to minimize the number of false positive predictions and ensure that the company only targets customers who are highly likely to book a holiday with British Airways.

Based on the evaluation results, we found that the XGBoost model achieved the highest AUC score among the three models. This indicates that the XGBoost model outperformed the others and demonstrated better predictive performance. It exhibited superior precision, recall, and F1-score, showcasing its ability to accurately classify positive and negative instances.

By achieving the highest AUC score and demonstrating better performance across multiple evaluation metrics, the XGBoost model proved to be the most suitable for predicting customer buying behavior in our study. We will further fine-tune this model to optimize its performance and make it more robust and reliable in real-world scenarios.

4.4 Model Tuning

Model tuning is an essential step in optimizing the performance of machine learning models. In this section, we applied hyperparameter tuning techniques to fine-tune the XGBoost model, which exhibited the best performance among the three models evaluated.

Hyperparameter tuning involves adjusting the model's hyperparameters to find the optimal configuration that maximizes performance. We used a random search approach to explore different combinations of hyperparameters efficiently. The used parameters are as follows:

Parameters	Values
max_depth	[10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110]
min_child_weight	[1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20]
gamma	[0.0,0.1,0.2,0.30000000000000004,0.4, 0.5,0.6000000000000001,0.7000000000000001,0.8,0.9,1.0]
tree_method	['auto', 'exact', 'approx', 'hist']
colsample_bytree	[0.0,0.1,0.2,...,0.5,000001,0.8,0.9,1.0]
eta	[0.0,0.010101010101010102,0.020202020202020204,.....,0.989898989,1.0]
lambda	[0.0,0.1,0.2,0.30000000000000004,0.4,0.5,0.6000000000000001,0.7000000000000001,0.8,0.9,1.0]
	[0.0,0.1,0.2,0.30000000000000004,0.4,0

alpha	.5,0.60000000000000001,0.70000000000000001,0.8,0.9,1.0]

Table 10 Hyperparameters

The best parameters are:

Parameters	Best values
tree_method	exact
min_child_weight	1
max_depth	80
lambda	0.1
gamma	1.0
eta	0.050505
colsample_bytree	0.6
alpha	0.3

Table 11 Best parameters

After performing hyperparameter tuning for the XGBoost model, we evaluated its performance using various metrics, including precision, recall, F1-score, and accuracy. The results after hyperparameter tuning for the XGBoost model were as follows:

Class	Precision	Recall	F1-score
0	0.88	0.95	0.91
1	0.95	0.87	0.91

Table 12 Results after tuning

By fine-tuning the hyperparameters, we were able to enhance the model's ability to capture complex patterns and make more accurate predictions. Through the process of hyperparameter tuning, we aimed to strike a balance between bias and variance, ensuring the model's

generalization to unseen data. The tuned XGBoost model exhibited improved precision, recall, F1-score, and accuracy, indicating its enhanced performance compared to the default configuration.

By optimizing the XGBoost model through hyperparameter tuning, we have further improved its predictive power and suitability for predicting customer buying behavior. The fine-tuned model can now be utilized for real-world applications, providing valuable insights and aiding decision-making processes.

4.5 Summary

In this chapter, we successfully applied three machine learning models to predict customer buying behavior. The XGBoost model, after tuning, demonstrated the highest performance, achieving an accuracy of 91% and precision of 95% outperforming the other models. This indicates that the XGBoost model is the most suitable for predicting customer buying behavior in our dataset.

Overall, the modeling phase of our study contributes valuable insights into understanding and predicting customer behavior, laying the foundation for the final chapter where we discuss the results, implications, and conclusions of our research.

5. Chapter Five: Results and Discussion.

5.1 Introduction

In this chapter, we present the results of our study, including the performance metrics and insights gained from the modeling process. We analyze the predictions made by the selected models and provide an evaluation of their effectiveness in predicting customer buying behavior. Additionally, we discuss the implications of the results and their potential impact on business decision-making.

5.2 Models Performance

We begin by examining the performance of the models in terms of various evaluation metrics. These metrics include precision, recall, F1-score, and accuracy. Each model was assessed individually to determine its ability to accurately classify customers as potential buyers.

5.1.1 Logistic Regression

The Logistic Regression model served as our baseline model. Its performance was evaluated based on precision, recall, F1-score, and support. The results showed a precision of 0.76 for classifying positive instances and a recall of 0.77, indicating that the model correctly identified a significant proportion of customers who were likely to make a purchase.

5.1.2 XGBoost Classifier

The XGBoost Classifier outperformed the other models in terms of precision, recall, and F1 score. It achieved a precision of 0.92 for positive instances and a recall of 0.86, demonstrating its ability to accurately identify potential buyers. The high F1 score indicated a good balance between precision and recall.

5.1.3 Catboost Classifier

The Catboost Classifier also exhibited strong performance, with a precision of 0.91 for positive instances and a recall of 0.86. These results indicated that the model effectively classified customers who were likely to make a purchase. The F1 score further confirmed the model's ability to strike a balance between precision and recall.

5.1.4 Tuned XGBoost Classifier

After performing hyperparameter tuning, the XGBoost Classifier demonstrated even better performance, surpassing the other models in terms of precision, recall, and F1 score. It achieved a precision of 0.95 for positive instances and a recall of 0.87 These results indicated that the model effectively classified customers who were likely to make a purchase.

5.3 Models Comparison

To compare the performance of the models more comprehensively, we assessed their AUC scores, which measure the models' ability to rank instances correctly. The XGBoost Classifier achieved an AUC score of 0.961, while the Catboost Classifier attained a slightly lower score of 0.960. These scores confirmed that both models were effective in distinguishing between potential buyers and non-buyers.

5.4 Business Implications

The results of our study have significant implications for business decision-making. The high precision scores achieved by the XGBoost indicate that this model is capable of accurately identifying potential buyers. This can help businesses optimize their marketing strategies by targeting specific customers who are more likely to make a purchase, thereby improving conversion rates and maximizing revenue.

Furthermore, the insights gained from the model predictions can inform personalized marketing campaigns, enabling businesses to tailor their offers and promotions based on individual customer preferences and behaviors. This targeted approach can enhance customer satisfaction and loyalty.

5.5 Limitations and Future Research

While our final model yielded promising results, it is essential to acknowledge its limitations. The models were trained and evaluated on a specific dataset, which may not fully represent all possible scenarios. Therefore, caution should be exercised when applying the models to different contexts or datasets.

In future research, it would be beneficial to explore additional features and data sources that could enhance the predictive power of the models. Additionally, investigating alternative modeling techniques and algorithms may provide further insights and improve the accuracy of predictions.

Overall, the results obtained from our study demonstrate the potential of machine learning models in predicting customer buying behavior. By leveraging these insights, businesses can make informed decisions, optimize their marketing strategies, and improve their overall performance in the market.

6. Chapter Six: Conclusion and Summary

6.1 Introduction

In this final chapter, we will provide a concise summary of the entire term paper, highlighting the main research question, objectives, methodology, results, and conclusions. We will emphasize the significance of the study and its implications for the field, discuss the broader implications of the findings, and highlight potential impacts on industry practices and future research directions.

6.2 Recap of Research Question and Objectives

The research question of this study was to predict customer booking behavior with British Airways using machine learning algorithms. The main objectives were to analyze the effectiveness of different algorithms, evaluate their performance using appropriate metrics, and provide insights for optimizing marketing strategies.

6.3 Methodology Overview

To address the research question, we employed three machine learning algorithms: Logistic Regression, XGBoost, and Catboost. The models were trained using balanced data and evaluated using various metrics, including AUC, recall, and precision. The performance of the models was assessed using cross-validation techniques, and hyperparameter tuning was applied to enhance their predictive power.

6.4 Summary of Results

Based on the evaluation metrics, the XGBoost algorithm outperformed the other models, exhibiting high precision, recall, and accuracy. The tuned XGBoost model achieved an accuracy of approximately 91% on the labeled data, with a precision of 95% for booking customers (class 1) and 88% for non-booking customers (class 0). The recall rates were 87% for booking customers and 95% for non-booking customers, indicating the model's ability to correctly identify these customer groups.

6.5 Conclusions

The study demonstrated the effectiveness of machine learning algorithms in predicting customer booking behavior with British Airways. The tuned XGBoost model provided valuable insights and recommendations for optimizing marketing strategies. By leveraging these insights, businesses can improve their targeting efforts, maximize conversion rates, and enhance revenue.

6.6 Significance and Implications

The findings of this study have significant implications for the field of customer behavior prediction and marketing strategy optimization. The use of machine learning algorithms can enable businesses to make informed decisions and allocate resources effectively to customers who are highly likely to book a service or make a purchase.

6.7 Broader Implications and Future Research Directions

The broader implications of this study extend to various industries that rely on customer behavior prediction and targeted marketing. The insights gained from this research can inform industry practices and guide decision-making processes. Future research can focus on exploring additional features like demographic information of customers and data sources to enhance the predictive power of the models. Investigating alternative modeling techniques and algorithms may also provide further insights and improve prediction accuracy.

In conclusion, this term paper addressed the research question of predicting customer booking behavior with British Airways using machine learning algorithms. Through a comprehensive methodology and evaluation process, the tuned XGBoost model emerged as the top-performing model, providing valuable insights for marketing strategy optimization. The study's significance lies in its implications for industry practices and future research directions, demonstrating the potential of machine learning in predicting customer behavior and improving marketing outcomes.

References

[1] Harsh Valecha, Aparna Varma, Ishita Khare, Aakash Sachdeva, and Mukta Goyal. Prediction of consumer behavior using random forest algorithm. In 2018 5th IEEE Uttar Pradesh section international conference on electrical, electronics, and computer engineering (UPCON), pages 1–6. IEEE, 2018.

[2] Neha Chaudhuri, Gaurav Gupta, Vallurupalli Vamsi, and Indranil Bose. On the platform but will they buy? predicting customers' purchase behavior using deep learning. Decision Support Systems, 149:113622, 2021.1

[3] Kiran Chaudhary, Mansaf Alam, Mabrook S Al-Rakhami, and Abdu Gu-maei. Machine learning-based mathematical modelling for prediction of so-cial media consumer behavior using big data analytics. Journal of Big Data,8(1):1–20, 2021.1

[4] Yoo, J. H., & Lee, S. J. (2019). Predicting customers' purchasing behavior using gradient boosting machine: Focusing on ancillary revenue in airline industry. Journal of Air Transport Management, 75, 97-107.

[5] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adap-tive synthetic sampling approach for imbalanced learning. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), pages 1322–1328. IEEE, 2008.1

[6] Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Ben-esty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. Noise reduction in speech processing, pages 1–4, 2009.1

[7] Alexander Kraskov, Harald St¨ogbauer, and Peter Grassberger. Estimating mutual information. Physical review E, 69(6):066138, 2004

[8] Cedric Seger. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing, 2018.1

[9] Emily A Boeke, Avram J Holmes, and Elizabeth A Phelps. Toward robust anxiety biomarkers: a machine learning approach in a large-scale sample. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 5(8):799–807, 2020.1