Predictive Modeling of Customer Bookings

Exploratory Data Analysis and Machine Learning

Agenda

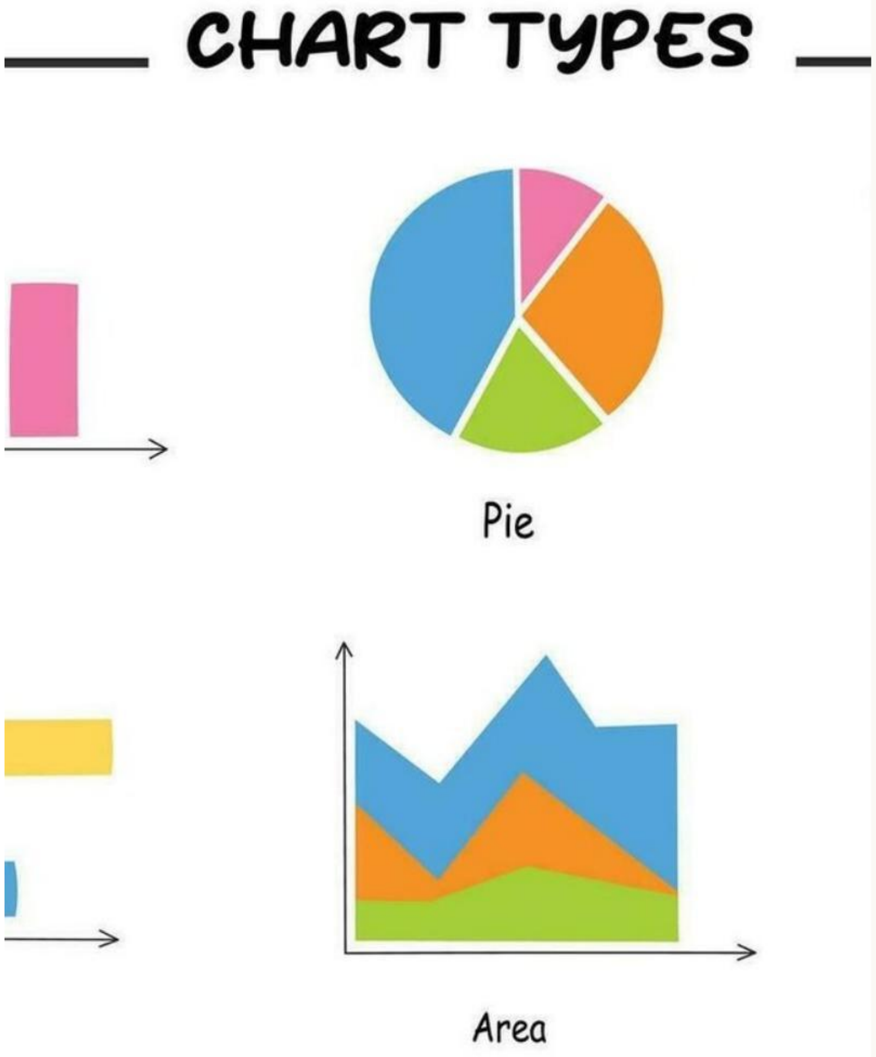
1. Introduction
2. Data Exploration
3. Data Preprocessing
4. Model Training
5. Model Evaluation
6. Feature Importance
7. Conclusion
8. Q&A

Introduction

Objective of the Study To predict customer booking completions using machine learning.

Exploration Purpose Analyze dataset to understand customer behavior and booking patterns.

Dataset Overview Contains 50,000 entries with 14 informative columns and no missing values.



Data Exploration

SlidesWizard.io

Data Types Overview The dataset contains 8 integer columns, 1 float column, and 5 object columns.

Flight Day Conversion Converted categorical 'Flight_day' to numerical format: Mon=1, Tue=2, etc.

Descriptive Statistics Mean values for key metrics: Purchase lead time is 84.94 days, Length of stay is 23.04 days, Flight hour averages to 9.07, and Flight duration is about 5.41 hours.

Data Preprocessing

Split Data into Features and Target

Identify and separate independent variables (features) from the dependent variable (target).

Categorical and Numerical Columns

Determine which columns hold categorical data (e.g., type of booking) and which are numerical (e.g., price).

Use OneHotEncoder for Categorical Data

Transform categorical columns into a numerical format that can be understood by machine learning algorithms.

Create Preprocessing Pipeline

Utilize ColumnTransformer to apply the necessary transformations to different columns in a streamlined manner.

Model Training

Model Used Random Forest Classifier with n_estimators=50 and random_state=42.

Training Dataset Model is trained on 80% of the available data.

Model Evaluation Cross-validation performed using 3 folds for better accuracy.

Model Evaluation

Precision

Precision for Class 0: 0.87, Class 1: 0.51 - Indicates the model's ability to reduce false positives.

Recall

Recall for Class 0: 0.98, Class 1: 0.14 - Reflects the model's ability to capture true positives.

F1-Score and Accuracy

F1-Score for Class 0: 0.92, Class 1: 0.22; Overall Accuracy: 0.85 - Balances precision and recall.

ROC AUC Score

ROC AUC Score: 0.78 for the Test Set; Mean Cross-Validated AUC: 0.612 - Measures overall model performance.

Feature Importance

Key Features Affecting Bookings

1. Purchase lead time: The advance booking timeline impacts customer choices.

Flight Details

2. Flight hour: Time of day influences passenger behavior.

Travel Preferences

3. Meal & seat preferences: Reflects customer expectations.

Booking Origins

4. Origin countries: Malaysia, Australia, Indonesia play crucial roles.

Conclusion

Model Performance

Random Forest model achieved 85% accuracy and 0.78 ROC AUC score.

Key Influencing Features

Important features include purchase lead time, flight hour, and length of stay.

Next Steps

Further hyperparameter tuning and exploring models like Gradient Boosting and XGBoost.

Data Imbalance Solutions

Investigate class imbalance and apply techniques such as SMOTE.