Athens University of Economics and Business

Statistical Machine Learning-Project 1

Tsadimas Anargyros

AM:f3612318

Supervisor:

D.Karlis

Department of Statistics

April 23, 2024

Contents

Abstract **iii**

Introduction **1**

1 Descriptive Analysis and Exploratory Data Analysis **2**

1.1 Data Origin and Collection 2

1.2 Variables Description 3

1.3 Data handling 3

1.4 Key Insights 4

2 Classification Mehods **6**

2.1 Logistic Regression 6

2.2 K-Nearest Neighbors 7

2.3 Linear Discriminant Analysis 8

2.4 Random Forrest 9

Bibliography **12**

List of Tables

1.1 Description of Variables in the Booking Dataset 3

2.1 Accuracy of Statistical Techniques 12

List of Figures

1.1 Histogram of Booking Lead Times. 4
1.2 Correlation Matrix for numeric variables 5
1.3 Histogram of average prices for hotel bookings 5
1.4 Bar Plot of Room Type 5
1.5 Bar Plot of Market Segment Type 5

2.1 ROC Curve demonstrating a random model's predictive efficiency 7
2.2 Distribution of Accuracies Over 100 Simulations/different splits 9
2.3 Out of bag error rate. 10
2.4 Variable importance plot from the Random Forest model. 11

Abstract

This project examines hotel booking data to predict which reservations might get canceled. We analyze different booking details and apply four statistical techniques: logistic regression, k-nearest neighbors (KNN), random forest, and linear discriminant analysis (LDA) to predict cancellation events. We assess the accuracy of these models and refine them to improve their prediction quality. The most effective model is identified by comparing their performances. Our report clearly describes our methods and results, using easy-to-understand language and clear visuals like tables and graphs. The goal is to help hotels better foresee cancellations to improve their planning and service.

Introduction

In this report, we explore how to predict when hotel bookings might be canceled. Understanding this helps hotels prepare better and offer rooms that might otherwise go unused. We use four different math-based methods to find patterns in booking data that suggest a reservation will not hold. Our work compares these methods to find the most reliable one. This study aims to give hotel managers a tool to anticipate and manage cancellations effectively, ensuring they can make the most of their available rooms.

Chapter 1

Descriptive Analysis and Exploratory Data Analysis

1.1 Data Origin and Collection

The data for this analysis originates from a comprehensive dataset of hotel bookings. This dataset was assembled by a consortium of hotels to better understand trends in reservation cancellations. The data is collected from several medium to large-scale hotels located in urban and resort locations, providing a diverse range of booking scenarios. Each entry in the dataset corresponds to a unique booking, capturing details from the initial reservation through to the final booking status (cancelled or not cancelled). The data were extracted from the hotels' reservation systems, ensuring accurate and real-time entry of each reservation's characteristics and outcome.

1.2 Variables Description

Table 1.1: Description of Variables in the Booking Dataset

Variable Name	Description
Booking_ID	A unique identifier for each booking.
Number of adults	The total number of adults included in the booking.
Number of children	The total number of children included in the booking.
Number of weekend nights	Count of weekend nights included in the booking.
Number of week nights	Count of weeknights included in the booking.
Type of meal	The type of meal plan booked (e.g., bed and breakfast, all-inclusive).
Car parking space	Indicates whether a parking space was included with the booking.
Room type	The type of room selected (e.g., standard, suite, family room).
Lead time	The number of days between the booking date and the arrival date.
Market segment type	The market segment from which the booking originated (e.g., direct, cor
Repeated	Indicates whether the booking was made by a returning customer.
P-C (Previous Cancellations)	The number of previous bookings that were canceled by the customer.
P-not-C (Previous Not Cancelled)	The number of previous bookings not canceled by the customer.
Average price	The average price per night of the booking.
Special requests	Number of special requests made by the guest.
Date of reservation	The date on which the booking was made.
Booking status	Final status of the booking (canceled or not canceled).

1.3 Data handling

We processed a dataset containing hotel booking details to prepare it for analysis. We cleaned the data by converting variables into numbers or factors wherever needed and categorizing and merging similar information together to simplify the models we plan to build.I also excluded variables that did not add information to the data,like the Booking ID variable and P-C, P-not-C,because they had almost all the observations in one level and were not statistical significant variables. Some dates in the dataset, specifically the 29th of February in non-leap years, were not valid and thus

excluded to maintain accuracy. We also handled missing values and standardized the format of dates, ensuring our data is consistent and reliable for predicting booking cancellations.

1.4 Key Insights

The preliminary analysis showed that a significant portion of bookings was made with a short lead time, suggesting that last-minute travel plans are common (Figure 1.1). Furthermore, special requests were found to correlate negatively with cancellations, implying that personalized service might reduce the likelihood of cancellation (Figure 1.2).

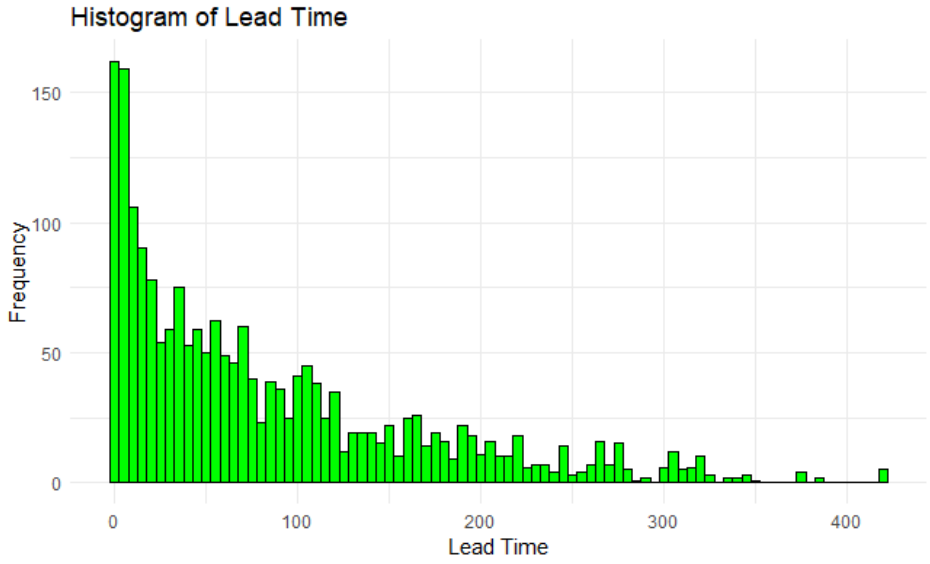

Fig. 1.1: Histogram of Booking Lead Times.

The correlation matrix (Figure 1.2) visually represents the relationships between the various numeric variables in the dataset. A correlation matrix is a useful tool in exploratory data analysis to quickly identify potential associations or dependencies between variables.

The histogram (Figure 1.3) shows the distribution of average prices for hotel bookings, mostly concentrated around a mean value, indicating a standard room cost. The decline in frequency for higher prices suggests that such bookings are less common.

These bar plots (Figures 1.4 and 1.5) display the frequency of bookings for different room types and market segments. The disparity in counts led to the consolidation of certain categories, simplifying the data for a clearer analysis and more robust modeling.

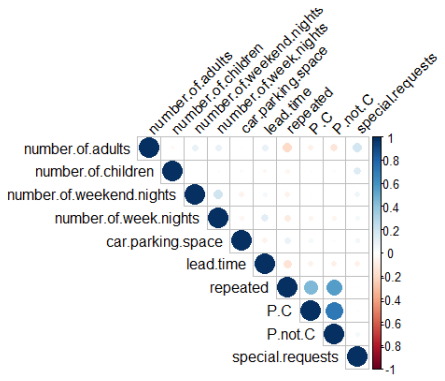

Fig. 1.2: Correlation Matrix for numeric variables

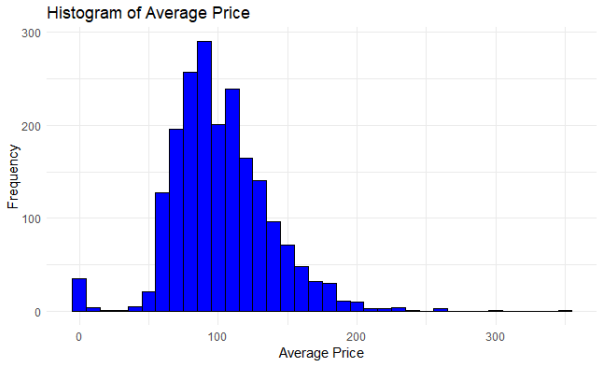

Fig. 1.3: Histogram of average prices for hotel bookings

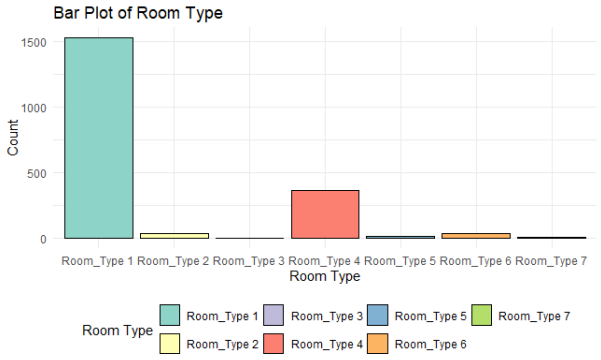

Fig. 1.4: Bar Plot of Room Type

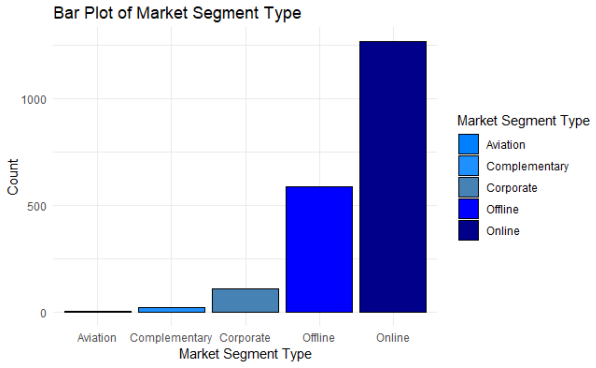

Fig. 1.5: Bar Plot of Market Segment Type

Chapter 2

Classification Mehods

2.1 Logistic Regression

The first method we used is logistic regression.Logistic regression predicts the probability of an event by fitting data to a logistic curve. It's used because it's great for binary outcomes like 'yes or no' decisions—perfect for understanding factors that influence customer bookings.

First we split our data into two distinct sets: a training set, which comprised 80% of the data, and a test set, the remaining 20%. The training set will help us build our predictive model, while the test set will be used to validate our predictions.

We fit a binomial Generalized Linear Model (GLM), which is a way to employ logistic regression considering the binary nature of our outcome – whether or not a booking was made. To further refine our model, we employed a technique called stepwise selection.Stepwise selection chisels away less significant variables to highlight the ones that truly impact the booking decision.

Once we had our predictive model, we then unleashed it upon the test set. Our goal here was to measure how accurately our model could predict whether a customer would make a booking.

To ensure the sturdiness and reliability of our model, we replicated this entire process 100 times(known as Bootstrap method), each time with a fresh split of data. This simulation approach buffers us against anomalies – ensuring our findings are robust.

The result is the model demonstrated an accuracy of about 80

This chart (see Figure 2.1) , called an ROC Curve.The ROC Curve is a plot that assesses the model's predictive efficiency.The AUC, or Area Under the Curve score, is 0.86 out of 1—showing

us our model is a strong predictor.

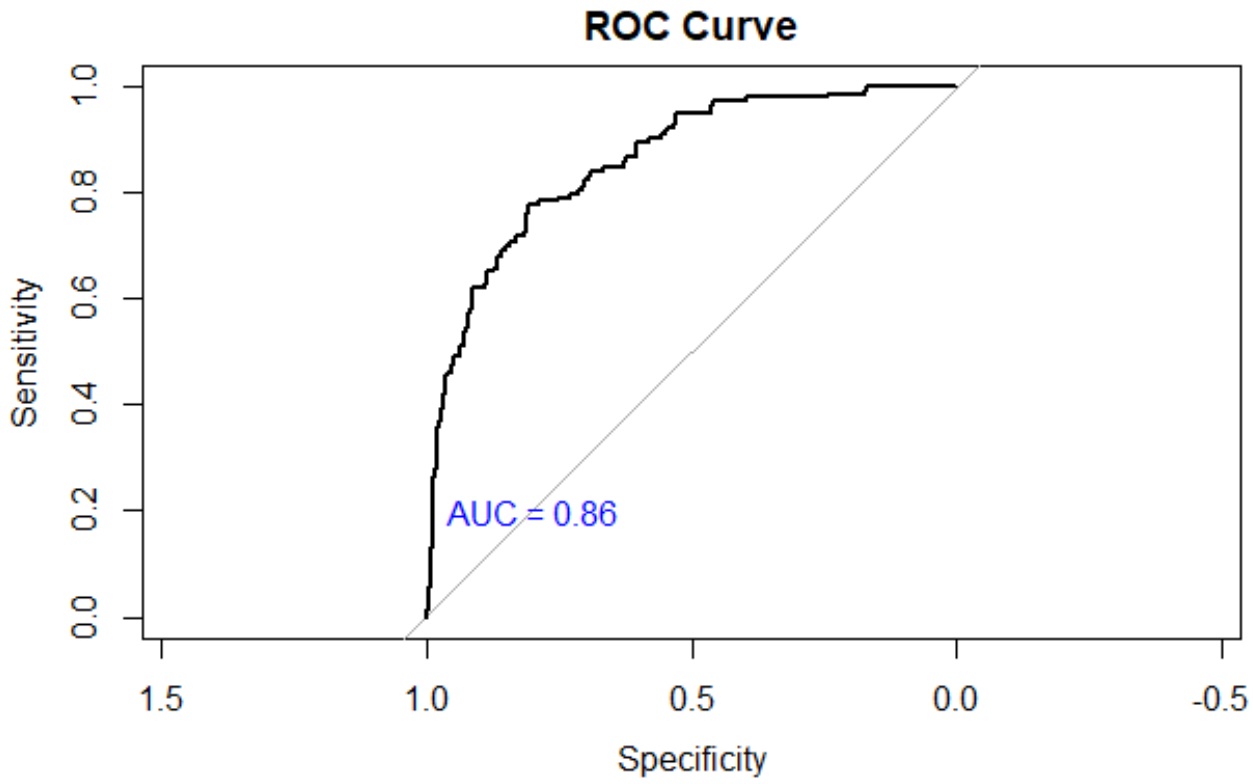

Fig. 2.1: ROC Curve demonstrating a random model's predictive efficiency

In conclusion, our logistic regression model, with its stepwise-refined variables and verified by repeated testing, stands as a robust tool.We will later see how it fares against the other methods.

2.2 K-Nearest Neighbors

The second method we used is K-nearest neighbors (KNN). KNN is a simple, intuitive method that predicts an item's classification based on the majority of its nearest neighbors.

In our analysis, we took our data and transformed all the categories into numbers because KNN relies on calculating distances to make predictions. Ensuring that all data points are numeric is essential since KNN determines the closeness of instances using distance metrics. After converting these categories, we then standardized all our data to the same scale. This step is crucial because it prevents any single feature from exerting undue influence over the prediction process due to

differences in measurement units or scale.

Additionally, we identified and removed any columns that were not contributing to predictive accuracy. This includes variables that were redundant, had little variance, or were not correlated with the outcome variable. Removing these non-contributive features simplifies the model, enhancing both its efficiency and performance by focusing only on relevant inputs.

Next, we mixed our data well and split it into two parts: 75% to learn from and 25% to test on.Using KNN, we looked at the 5 closest neighbors to guess if a customer would make a booking.

Choosing 5 neighbors offers a balance between noise reduction and generalization. With too few neighbors, the model becomes sensitive to outliers, where a single unusual neighbor can sway the prediction inaccurately. Conversely, too many neighbors make the classification boundary too general, possibly ignoring smaller yet crucial patterns in the data, which could be significant for specific cases like predicting customer behavior in booking scenarios. Furthermore, an odd number helps avoid ties in voting, ensuring a clear majority. Testing different values for 'k' can sometimes result in better performance, but in this case, 'k=5' provided a reasonable trade-off between accuracy and computational efficiency, making it a practical choice for our dataset size and diversity.

The accuracy we got from this approach is 74%, which is a good start, especially considering how straightforward the method is.

2.3 Linear Discriminant Analysis

The third method used is Linear Discriminant Analysis (LDA).LDA is a statistical method used for classification.It predicts categories by finding a linear combination of features that separates them most distinctly.

For our project, we first made sure all our data, like customer details, was in numbers because LDA needs numbers to work with. We also made sure no single type of data could overpower the others by standardizing everything, which means adjusting the data so each type has an equal chance to influence the results. We then split our data, using 75% to train our model and 25% for testing. This split was done over 100 trials to make sure our results were solid. We trained our model with the data, but we didn't include the booking status; we wanted to predict this. After training, we used our model to predict booking statuses and checked how well it did by comparing

the predictions to the actual results. This comparison helps us understand the model's accuracy.

The whole process showed our model was 79,74% accurate, meaning it did a good job at predicting if a customer would book based on the patterns it learned from the data.(see Figure 2.2).

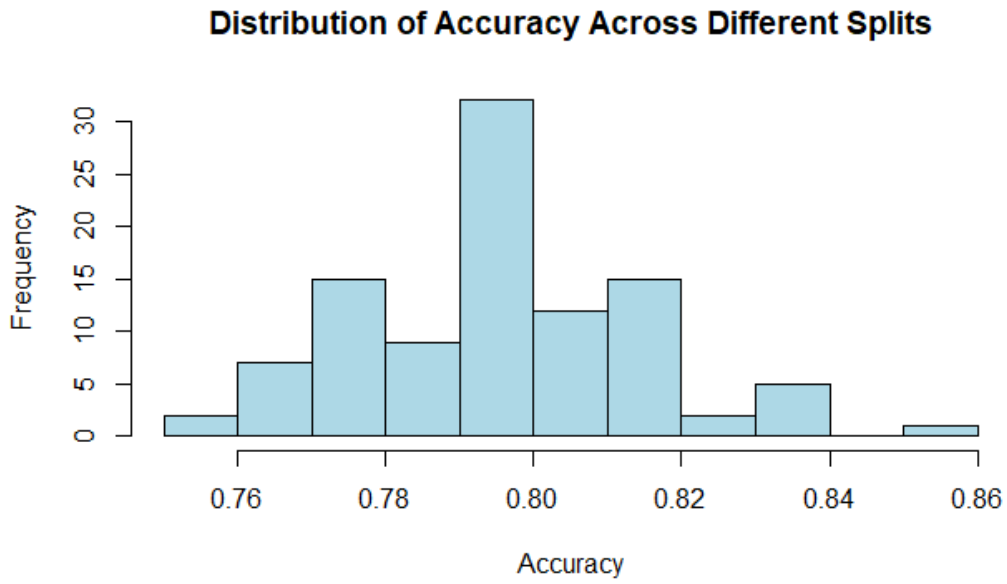

Fig. 2.2: Distribution of Accuracies Over 100 Simulations/different splits

2.4 Random Forrest

The fourth method we used is Random Forest.This is a powerful method to make predictions or classify outcomes based on various input variables. It builds on the idea of decision trees, which are simple models that make decisions based on answering a series of questions about the features of a dataset. Random Forest combines many of these trees to improve the accuracy and stability of predictions.

In our project to predict booking status, we began by ensuring our target variable, 'booking.status', was formatted correctly as a categorical factor, since Random Forest works best with categorical targets for classification tasks.Our Random Forest model used all available data, built 200 trees, and allowed three variables to be considered at each decision point within the trees. We also calculated the importance of each variable to understand their contributions to prediction

accuracy.

After building the model, we visualized the error reduction as more trees were added, helping us understand when adding more trees does not significantly improve accuracy. We also evaluated the model's out-of-bag (OOB) error rate, an in-built method for estimating the model's error on unseen data.(Figure 2.3) This suggests that beyond this point,50 trees, adding more trees does not significantly improve the model's performance.

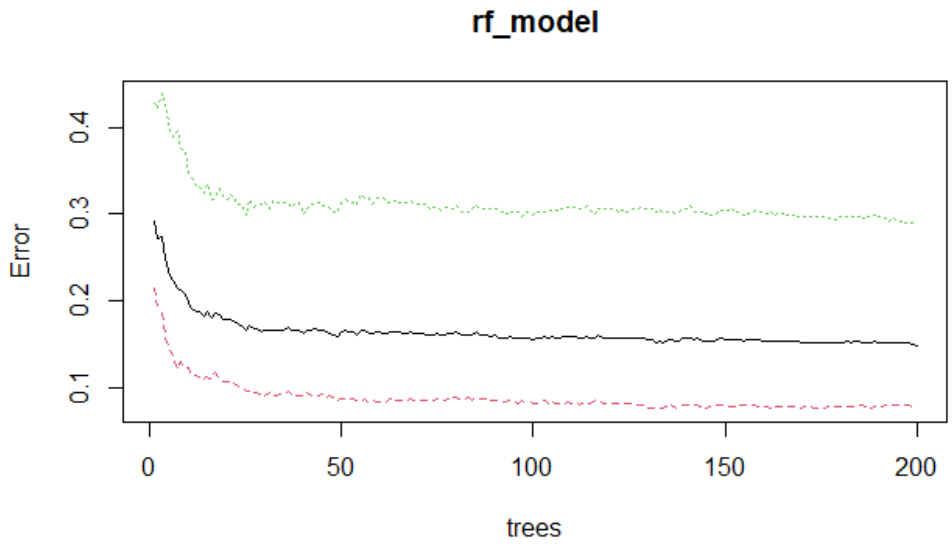

Fig. 2.3: Out of bag error rate.

To ensure optimal model performance, we experimented with different numbers of trees and variables at each split, finding the best settings that minimized prediction errors (50,3).Then using this setting we simulated the process 100 times. The final model showed an ability to predict booking statuses with about 84.2% accuracy.

A variable importance plot (Figure 2.4) revealed which factors were most influential, providing insights into features that might need more focus for strategic decisions or improvements.As shown in Figure 2.4, the variable importance plot indicates that 'lead.time' is the most critical factor influencing the model's predictions.

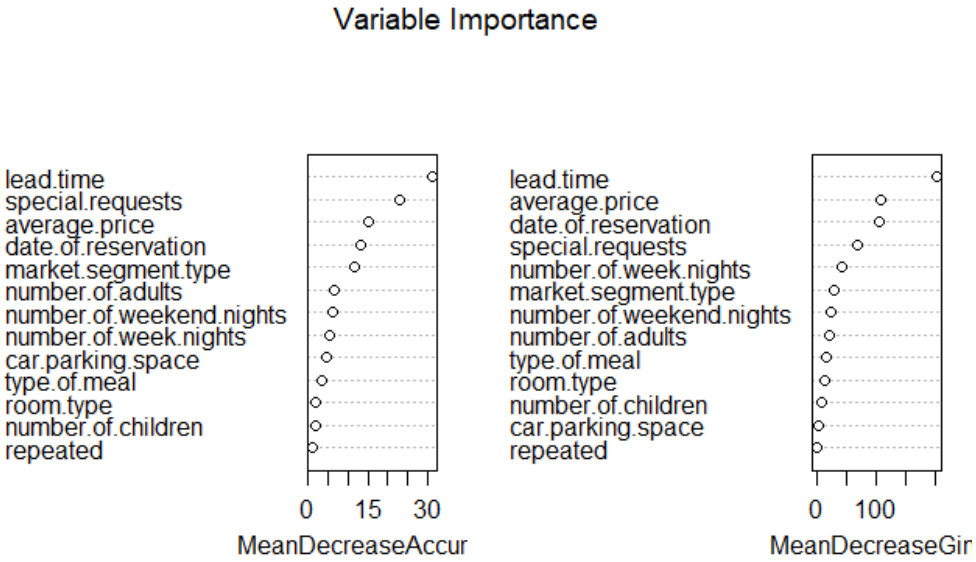

Fig. 2.4: Variable importance plot from the Random Forest model.

Conclusions

In conclusion, this report has meticulously explored the application of various statistical techniques to predict hotel booking cancellations. Each method—logistic regression, k-nearest neighbors (KNN), linear discriminant analysis (LDA), and Random Forest—offers unique insights into the factors that may lead to a reservation being canceled.

Table 2.1: Accuracy of Statistical Techniques

Method	Accuracy (%)
Logistic Regression	80.4
K-Nearest Neighbors (KNN)	73.9
Linear Discriminant Analysis (LDA)	79.74
Random Forest	84.2

Our comprehensive analysis found that Random Forest emerged as the superior technique, with an impressive 84.2% accuracy in predicting cancellations(Table 2.1) . This method's strength lies in its ensemble approach, leveraging multiple decision trees to reach a consensus that outperforms any single model. It effectively identified 'lead time' as the most influential predictor, which aligns with the intuitive notion that bookings made well in advance have a higher likelihood of cancellation.

The project offers a robust framework for anticipating cancellations. By incorporating these predictive models into their operations, hotel managers can better manage their inventory and design targeted strategies to mitigate potential losses from cancellations. Future studies could further refine these models, explore additional predictors, and possibly integrate real-time data to enhance predictive accuracy.

Bibliography

[1] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer Texts in Statistics, New York, NY: Springer, 1st edition, 2013.

[2] Dimitris Karlis, *Statistical Machine Learning Lecture Notes*, Department of Statistics, Athens University of Economics, karlis@aueb.gr, Athens, Feb, 2024.