Hotel Booking Cancellation Prediction

AUGUST 25, 2025

Course Title: Machine Learning for Business
Instructor's Name: Professor Diksha Godbole
Authored by: Nahid Aalam [1039075]

1. Problem Statement

One of the most significant challenges hotels face today is **booking cancellation**, which can affect hotel operations, revenue forecasts, and customer experience in various ways.

From a financial perspective, cancellations can affect revenue forecasting as hotels depend on advance bookings to predict their income. Cancellations, especially last-minute ones, lead to discrepancies between projected and actual revenue, resulting in lost income opportunities.

Cancellations also hinder and impact operational planning. Hotels often overbook rooms, anticipating a certain percentage of cancellations. However, when actual cancellations deviate from expectations, it may lead to either wasted resources (empty or unused rooms, idle staff) or overbooking, which negatively affects customer satisfaction and operational efficiency.

Customer experience is another concern that may be influenced by booking cancellation. Overbooking or last-minute cancellations can result in dissatisfaction, causing potential loss of future loyal customers.

To address these issues, it is crucial to predict whether a booking will be canceled before the date of arrival. By accurately forecasting cancellations, hoteliers can support smarter decisions in room allocation, staffing, revenue management, and customer service, ultimately reducing financial loss and improving guest satisfaction.

2. Objectives

The main goal of this project is to build a predictive system based on a machine-learning algorithm to determine whether a hotel booking will be canceled accurately. This involves data preprocessing, model training, and evaluating using multiple classification models, including **Logistic Regression, Decision Trees, Random Forest,** and **Linear SVM**. The model performance will be assessed using standard classification metrics: **accuracy**, **precision**, **recall**, and **F1 score**. Special emphasis will be placed on the "*canceled*" class (class 1), as it is the main question of this project.

3. Dataset Overview

The dataset used for this project is sourced from Kaggle and is titled *hotel_bookings.csv*. It contains various booking records from two different types of hotels (a city hotel and a resort hotel) in Portugal, covering the period from 2015 to 2017. The dataset consists of 119,390 records and 36 columns. It contains both numerical and categorical features such as lead time, deposit type, meal type, customer type, number of special requests, length of stay, number of adults, children, and/or babies, and whether the booking was canceled.

A subset of key variables is listed below:

Index	**Variable**	**Description**
1	Hotel	Type of the hotel (Resort Hotel, City Hotel)
2	Cancellation_status	Reservation cancellation status (0 = not canceled, 1 = canceled)
3	Lead_time	Number of days between booking and arrival
4	Arrival_date_year	Year of arrival
5	Arrival_date_month	Month of arrival
6	Arrival_date_week_number	Week number of the year for arrival
7	Arrival_date_day_of_month	Day of the month of arrival
8	Weekend_nights	Number of weekend nights (Saturday and Sunday) the guest stayed or booked
9	Weeks_nights	Number of weeknights the guest stayed or booked
10	Adults	Number of adults
11	Children	Number of children
12	Babies	Number of babies
13	Meal	Type of meal booked (BB, FB, HB, SC, Undefined)
14	Country	Country of origin of the guest
15	Market_segment	Market segment designation

4. Data Preprocessing

Data preprocessing is a crucial step in any machine learning project. It involves cleaning and transforming raw data into a format that machine learning algorithms can effectively understand and use. In this step, the dataset is prepared for modeling by performing several crucial tasks. Some of these steps are described in detail below:

4.1. Feature Selection and Engineering

Feature engineering can help with reducing the dimensionality of the dataset, which improves computational efficiency, and model accuracy. It also reduces the risk of overfitting and improves the generalizability of the model

4.1.1. Remove Directly Related Features

Directly related features are those that are closely or directly tied to the target variable. These features may contain information that either duplicates or directly reveals the outcome that the model is trying to predict. Removing these types of features can help the algorithm avoid **data leakage** which could give the model unfair predictive advantages and lead to misleading evaluation results. For example, "*reservation status*" in this dataset is a feature that indicates the current status of a reservation, such as "*Canceled*"," *Check out*", and "*No show*". If the value is "*Canceled*" or "*No show*" it directly reveals that the booking was canceled (corresponding to a target value of 1 in cancellation status), disclosing the outcome the model is supposed to predict on its own.

4.1.2. Drop Irrelevant Features

Some features probably do not have specific meaningful information for the predictive model. They may add noise and increase computational complexity without improving model performance. These features can be considered **irrelevant** and should be removed from dataset. In this case, features such as "*Arrival date year*", "*assigned room type*" and some other categorical features like "*country*", "*agent*", "*company*", "*name*", "*email*", "*phone number*", and "*credit card*" are considered irrelevant, as they do not contribute meaningful predictive value.

4.2. Handle Missing Values

Missing values should be addressed based on the context of the analysis, and the questions the project aims to answer. The only column with the missing values in this dataset is the "*children*" column which has a very small number of missing values (4). Given its potential impact on the model, it is more appropriate to retain this feature. Analyzing the distribution of values in this column shows that the majority of bookings involve 0 children, which is a strong mode. Since the "*children*" column is a numerical variable, it would be reasonable to replace the missing values of this column with 0 (the mode of the column).

4.3. Handle Noisy Data

Noisy data refers to data that contains a lot of random errors, irrelevant information, or inconsistencies, which can negatively affect the model's ability to make accurate predictions. Features such as "*average daily rate (ADR)*", "*adults*", "*children*", and "*babies*" are some of the variables in the dataset which contain inconsistent values. Depending on the number of these errors, and the overall importance of the feature, the affected features are either removed or corrected to improve data quality.

4.4. Encode Categorical Variables

To prepare the dataset for training, several preprocessing steps were applied specifically to the categorical features:

The "*arrival date month*" feature, originally in string format, and representing an ordinal variable, was converted to numerical values using label encoding. For features with multiple values, such as “*hotel*”, “*meal*”, “*market segment*”, “*distribution channel*”, “*reserved room type*”, “*deposit type*” and “*customer type*” one-hot encoding was applied, since these are categorical variables without any natural ordering. Additionally, the target variable “*cancellation status*”, was label-encoded to transform it into a binary numeric format (0 for not canceled, 1 for canceled).

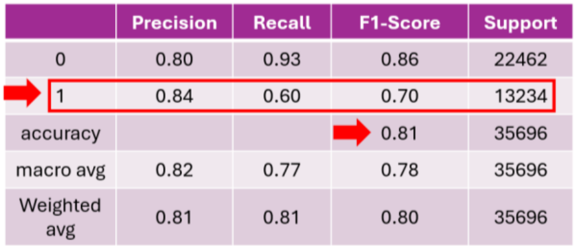
5. Model Building

Four machine-learning models were trained and evaluated using a test dataset.

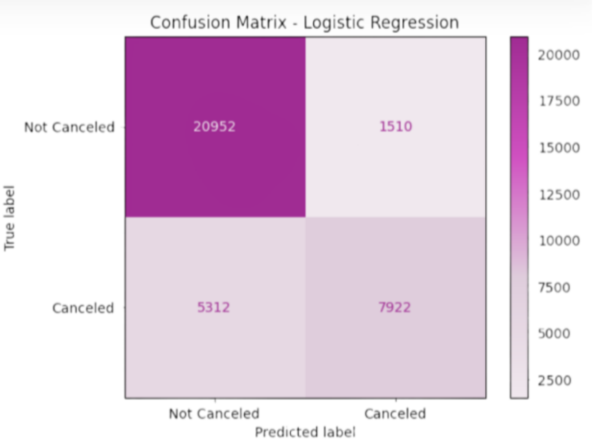
5.1. Logistic Regression

Our Logistic Regression model's performance is as follows:

The overall accuracy is 81%, which means the model is doing a solid job in predicting whether a booking will be canceled or not. The precision for cancellations (class 1) is 84%, indicating that 84% of the bookings predicted as canceled were actually canceled. However, the recall for class 1 is lower at 60%, meaning the model correctly identified only 60% of all actual cancellations.



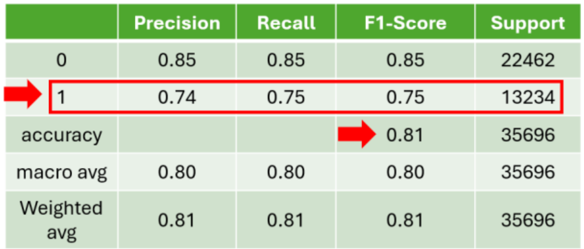
	Precision	Recall	F1-Score	Support
0	0.80	0.93	0.86	22462
1	0.84	0.60	0.70	13234
accuracy			0.81	35696
macro avg	0.82	0.77	0.78	35696
Weighted avg	0.81	0.81	0.80	35696



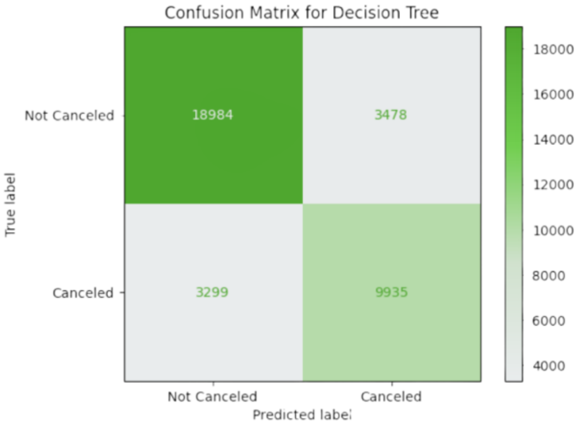
The F1-score for class 1 is 70%. This metric reflects the balance between precision and recall, and the lower score shows that there is a noticeable gap between those two metrics. This suggests the model is more cautious and tends to favor precision over recall in predicting cancellations. However, it ends up missing some cancellations.

5.2. Decision Tree

The Decision Tree model achieved an accuracy score of 81%, the same as the Logistic Regression model. The precision for class 1 is 74%, which means that 74% of the bookings that the model predicted as canceled were actually canceled. The recall for class 1 is 75%, which means that the model correctly identified about three out of every four actual cancellations. The F1-score for class 1 is also 75%, reflecting a good balance between precision and recall.



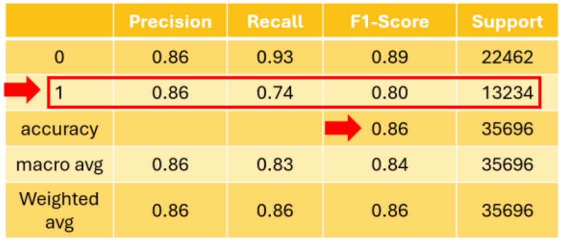
	Precision	Recall	F1-Score	Support
0	0.85	0.85	0.85	22462
1	0.74	0.75	0.75	13234
accuracy			0.81	35696
macro avg	0.80	0.80	0.80	35696
Weighted avg	0.81	0.81	0.81	35696



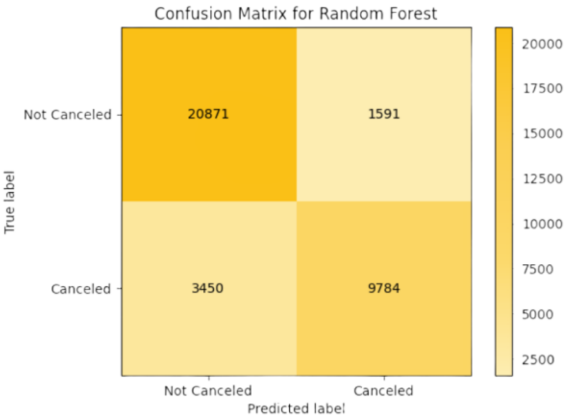
The confusion matrix shows that there are still some "*False Positives*" and "*False Negatives*", however, the model is doing a relatively good job of minimizing them. Overall, the Decision Tree model is performing well on the test data and seems to be a good model for predicting cancellations, though there is still room for improvement.

5.3. Random Forest

The Random Forest model delivered the strongest performance compared to the two previous models, with an accuracy of 86%. For class 1, precision is 86%, meaning the model is highly reliable when it predicts a booking will be canceled. Recall is 74%, which is better than Logistic Regression (60%) and slightly lower than Decision Tree (75%), but still strong. The F1-score is 80%, indicating a well-balanced tradeoff between precision and recall.



	Precision	Recall	F1-Score	Support
0	0.86	0.93	0.89	22462
1	0.86	0.74	0.80	13234
accuracy			0.86	35696
macro avg	0.86	0.83	0.84	35696
Weighted avg	0.86	0.86	0.86	35696



The confusion matrix shows this balanced performance, with fewer "*False Negatives*" and "*False Positives*" compared to the other models.

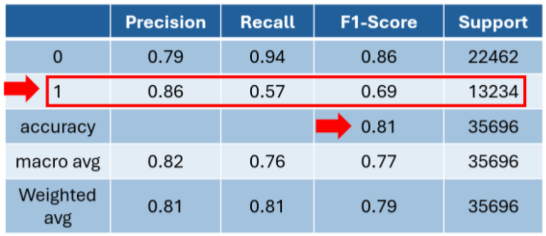
Overall, the Random Forest model stands out for its combination of accuracy, consistency across metrics, and improved performance in identifying both canceled and non-canceled bookings.

5.4. SVM

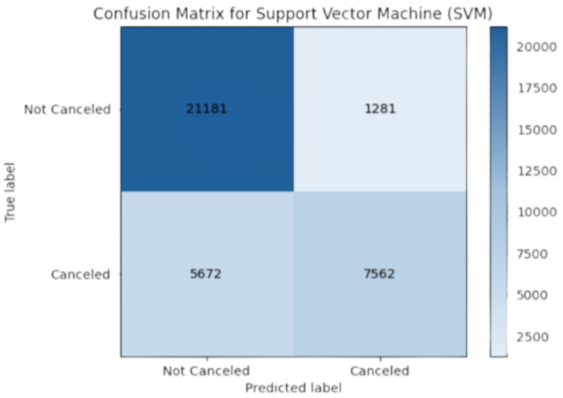
The Linear SVM model achieved an accuracy of 81%, matching the performance of both the Logistic Regression and Decision Tree models but falling short of the Random Forest's 86%. When it comes to predicting cancellations (class 1), the precision is 86%, the highest among all models, meaning when the Linear SVM predicts a booking will be canceled, it's usually right. However, the recall is just 57%, the lowest of all models, which shows that Linear SVM misses a significant number of actual cancellations. The F1-score is 69%, reflecting this imbalance between precision and recall.

The confusion matrix highlights this trade-off: while false positives are relatively low, false negatives (missed cancellations) are more common than in the other models.

Overall, the Linear SVM model is very confident in its cancellation predictions but tends to under-detect them. It’s strong on precision but not ideal if capturing as many cancellations as possible is a priority. For balanced performance, Random Forest remains the best option.

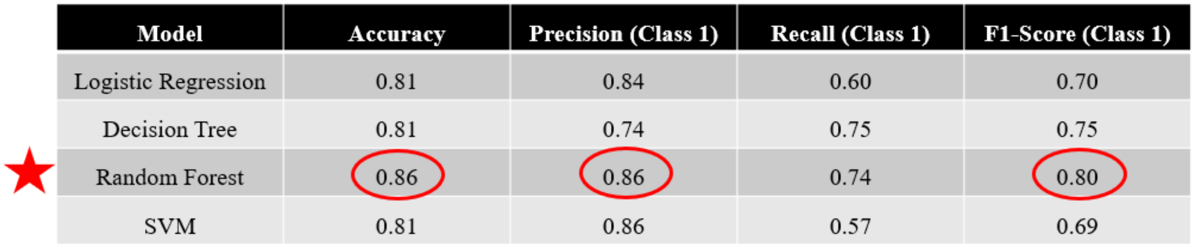


	Precision	Recall	F1-Score	Support
0	0.79	0.94	0.86	22462
1	0.86	0.57	0.69	13234
accuracy			0.81	35696
macro avg	0.82	0.76	0.77	35696
Weighted avg	0.81	0.81	0.79	35696



6. Conclusion

Among the four models evaluated, the Random Forest stands out as the top performer. It achieved the highest accuracy (86%) and F1-score for cancellations (80%), indicating a strong balance between precision and recall. While Linear SVM had the highest precision (86%), its low recall (57%) resulted in a lower F1-score (69%), meaning it missed many actual cancellations. Logistic Regression and Decision Tree showed acceptable performance, but Random Forest provided the most consistent and reliable results for predicting cancellations.



Model	Accuracy	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)
Logistic Regression	0.81	0.84	0.60	0.70
Decision Tree	0.81	0.74	0.75	0.75
★ Random Forest	0.86	0.86	0.74	0.80
SVM	0.81	0.86	0.57	0.69

The predicted cancellation rate based on the Random Forest model shows that around 32% of the bookings are expected to be canceled (roughly 1 in 3 bookings), which is relatively high. This high cancellation rate highlights a significant risk of lost revenue and operational inefficiencies, such as disruption to staff scheduling, inventory planning, and room preparation, if cancellations are not proactively managed.

The strong performance of the Random Forest model provides a reliable foundation for improving overbooking strategies, reducing financial risk, and enhancing guest satisfaction.

As a recommendation for further research, conducting a feature importance analysis would be a valuable next step. It can reveal which booking attributes have the greatest impact on cancellations, enabling hotel managers to make data-driven decisions and potentially reduce cancellation rates through policy adjustments, personalized communication, and strategic interventions aimed at reducing future cancellation rates. Additionally, the application of other SVM models (RBF and Polynomial kernels) alongside Linear SVM, would provide a comparative understanding of model performance. This could help identify the most suitable kernel function for accurately predicting booking cancellations, therefore enhancing the robustness and reliability of the predictive framework.