Hotel Reservation Cancellation Risk Model

Capstone 2 Final Report

By Cameron Hicks

Hotel reservation cancellations can have a significant impact both on top line revenue and bottom line profitability, particularly cancellations closer to the reservation arrival date which leave a limited window for replacement bookings. In this project I analyze a data set of reservations for two hotels in Portugal; one resort hotel located on the coast and one city hotel located in Lisbon. The objective of this project is to develop a machine learning model that predicts the probability of new reservations cancelling.

To accomplish this, two data sources were used:

- **Hotel Booking Demand** - provided by Kaggle - https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand/data
- **Open Meteo Weather API** - https://open-meteo.com/

The Hotel Booking Demand dataset contains the reservation-level data for the two hotels in Portugal. This data set contains over 119,000 reservation records and 32 attributes including cancellation status, booking lead time, arrival dates, market segments and more. The cancellation status attribute, titled ‘is_canceled’, was identified as the target data for our project of predicting cancellations.

While this dataset provides extensive reservation-level information, it lacked contextual variables that may influence traveler behavior. One of the largest variables that could impact hotel cancellations are weather forecasts and conditions. To enrich the existing hotel reservation data, I incorporated historical weather data from the Open-Meteo API.

The data wrangling process was conducted in a jupyter notebook using python and common python packages: Pandas, Numpy, Matplotlib.pyplot, seaborn, requests, and datetime. The Hotel Booking Demand dataset was imported from a csv file downloaded from the Kaggle datasource. Upon the initial inspection the hotel data contained 119,390 reservation records and 32 attributes for each record. Several attributes could be immediately identified as potential top features for the cancellation prediction model including Lead Time, Is Repeated Guest, Previous Cancellations, Customer Type, Reservation Status, and Reservation Status Date.

During inspection, missing values were identified in the children, country, agent, and company columns. These were handled as follows:

- Missing values in children were replaced with 0, under the determination that null value indicated no children traveling.
- Missing values in agent and company were also replaced with 0 indicating that the reservation was not associated with a travel agent or corporate account.
- Missing values in Country were replaced with the string “Unknown”.

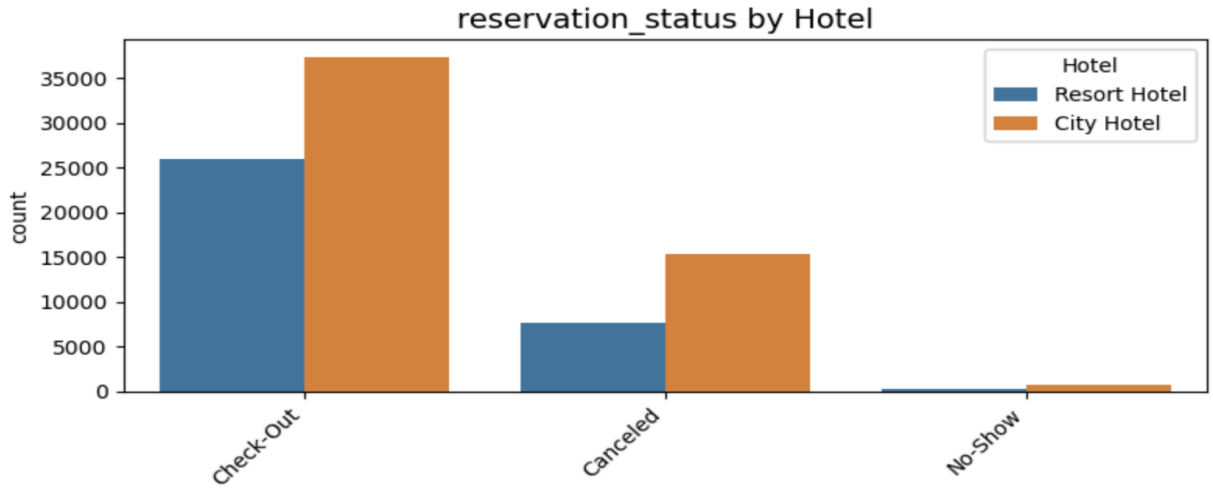
Once missing values were addressed, the dataframe contained 31,994 duplicate records, all were dropped to preserve the integrity of the model training. To prepare for pulling the weather data from the Open Meteo API a column titled “Hotel City” was added to the dataframe and a column titled “arrival date” was created that merged the data in the three separate date columns: ‘arrival date year’, ‘arrival date month’, and ‘arrival date day of month’. Then, to prepare the dataframe for feature engineering several of the attribute data types were changed. The attributes Children, Agent and Company were changed from float to integer, and all date attributes were converted to datatime data types.

To pull the desired weather data, I used the hotel coordinates which were provided by the datasource Kaggle, and created a list of unique arrival dates to minimize redundant API calls. Using these date points I generated the weather observations for each arrival date at each hotel. These weather observations included maximum temperature, minimum temperature, precipitation summary, and a weather code. Using information provided by Open Meteo, I created a mapping dictionary with the weather code descriptions to improve interpretability. The resulting weather dataset was validated to ensure no missing values were present and then merged with the hotel reservation dataset. The final combined dataset, referred to as **hotel_weather_df,** marked the conclusion of the data wrangling phase.

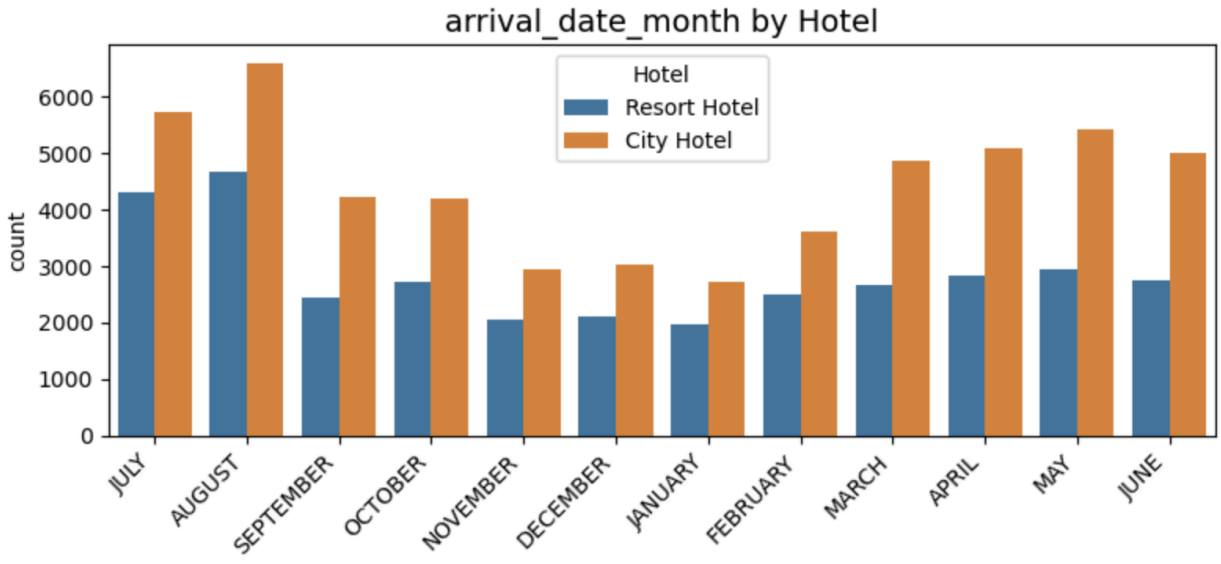
Exploratory data analysis was conducted in a jupyter notebook using python and common python packages: Pandas, NumPy, SciPy, Matplotlib, Seaborn, and Scikit-learn. After cleaning, dropping duplicates, and merging during the data wrangling step the final dataset contained 87,396 reservation records, of which 53,428 corresponded to the city hotel and 33,968 to the resort hotel.

This imbalance prompted several exploratory questions, including whether the city hotel experiences shorter lengths of stay, higher business travel volume, or fundamentally different cancellation behavior than the resort hotel. These questions ultimately led to a key modeling consideration: do the two hotel types exhibit sufficiently similar patterns to justify a single predictive model?

To explore this question I began with feature comparison between the two hotels. In the chart below we can see the comparison of the reservation status for each hotel. Since there are significantly more reservation records for the City Hotel vs the Resort hotel, it is expected that the chart will show a higher number of reservations for the City Hotel. The insight we can gain from this chart is that we can see the same ratio of reservations for each hotel in each reservation status.

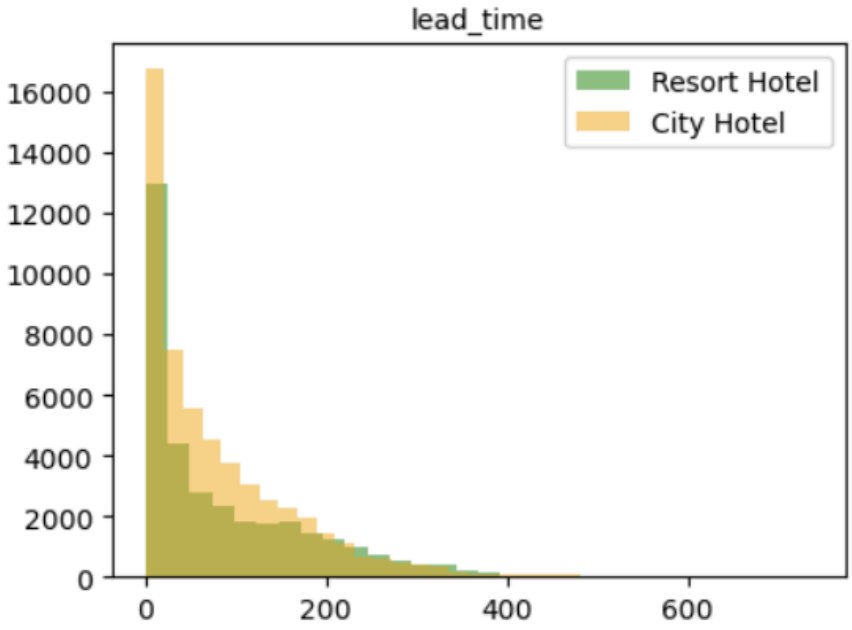


We see a similar pattern when inspecting the arrival date month by hotel chart.

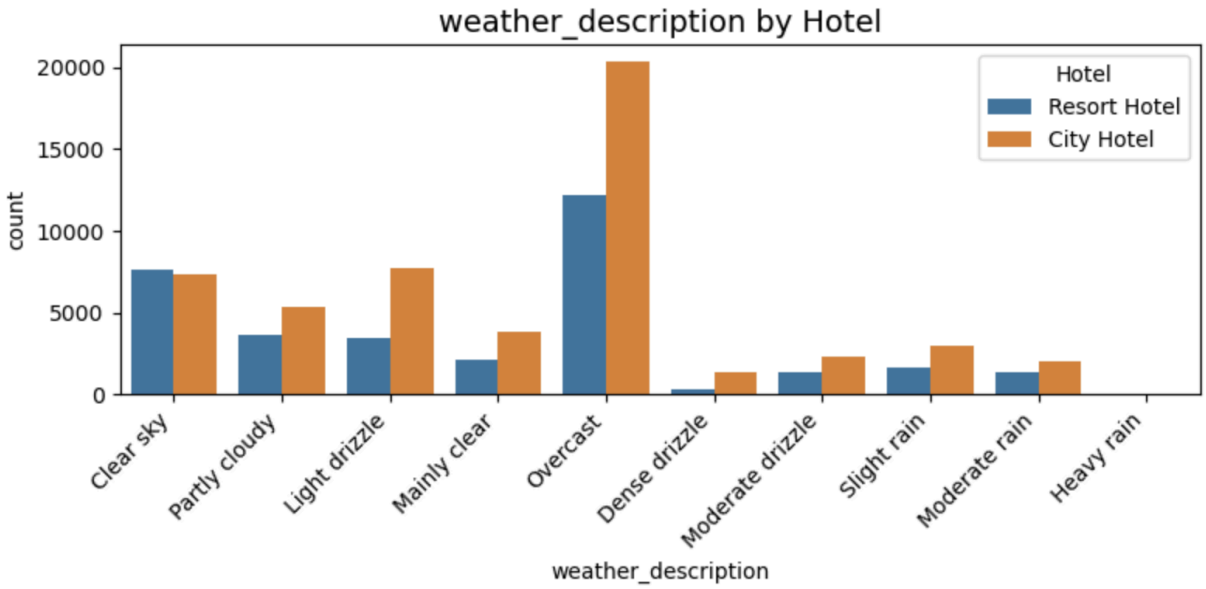


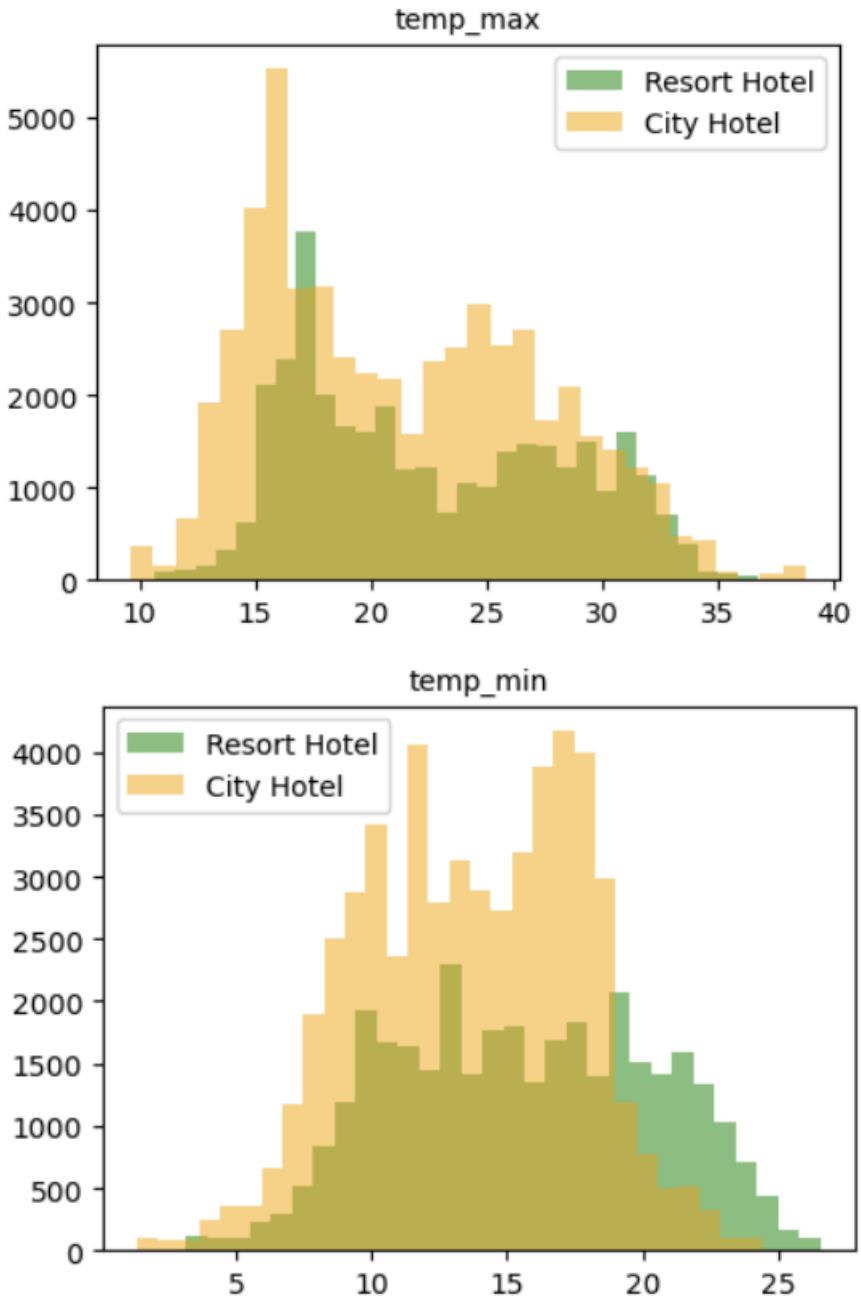
Furthermore, while inspecting the booking lead time it is clear that both hotels experience similar booking windows. We can also see that this chart is right skewed, indicating that the majority of reservations are made closer to the arrival date. This

insight shows the importance of an accurate cancellation prediction model to allow hotels enough time to prepare for cancellations.



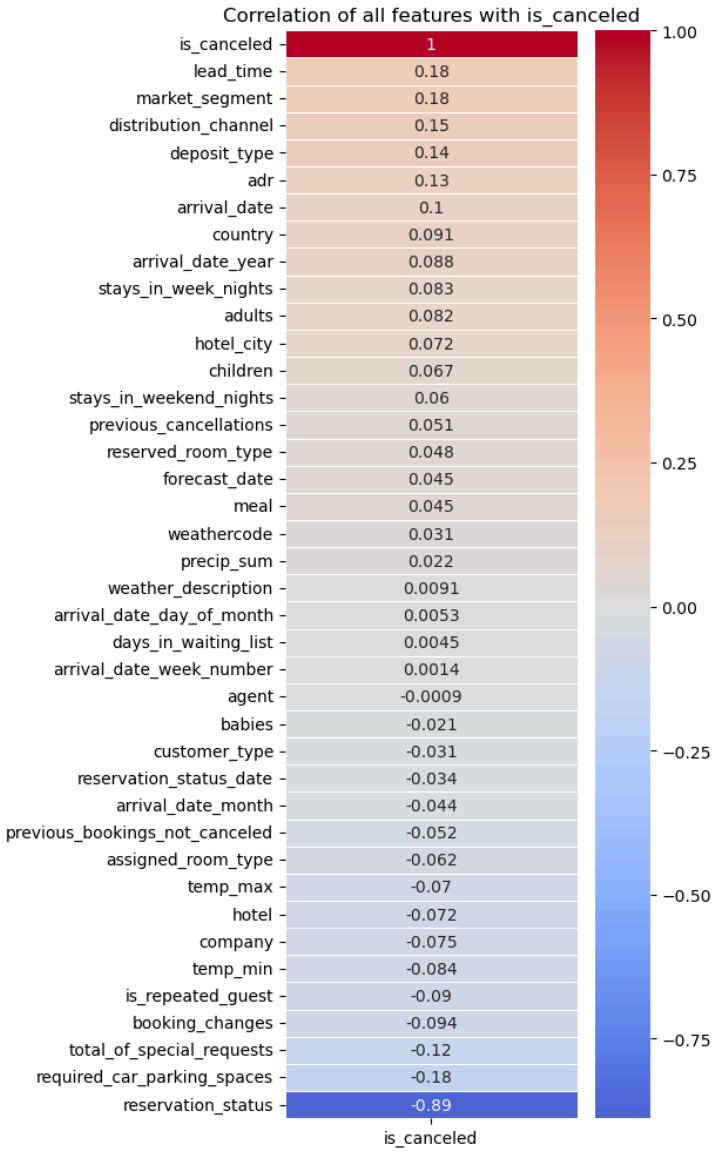
Finally, inspecting the weather data for each hotel shows the same pattern seen with the above.



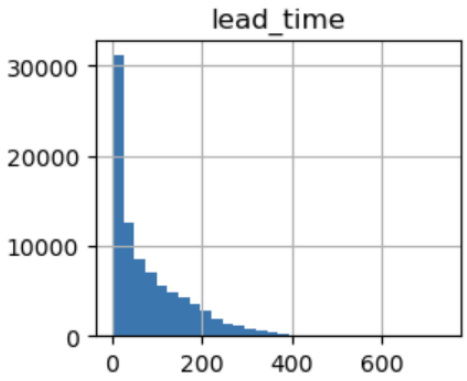


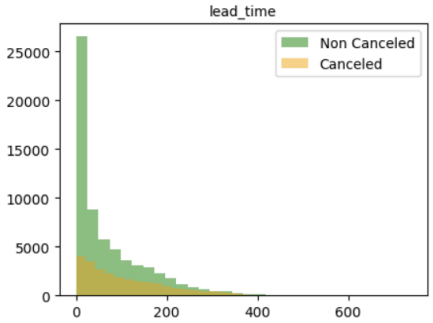
These insights strongly suggest that both hotels experience similar guest booking behavior, which supports using one model for both hotels.

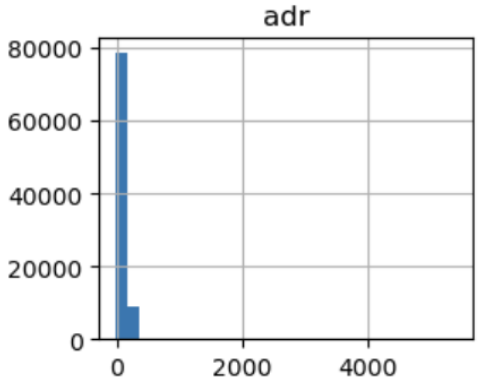
To identify features most strongly associated with cancellations, I generated a Pearson correlation heatmap for the dataset. The full correlation matrix was difficult to interpret visually, so I created a focused heatmap centered on the **is_canceled** feature. This chart revealed the most influential predictors.

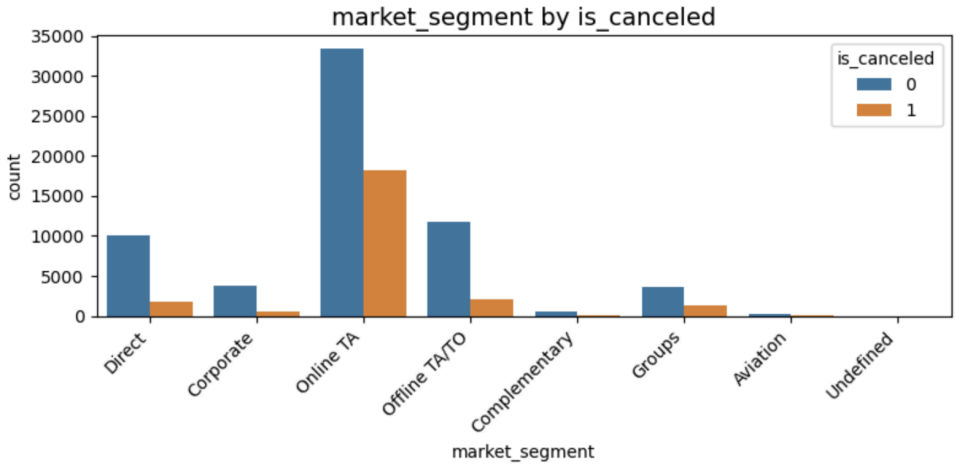


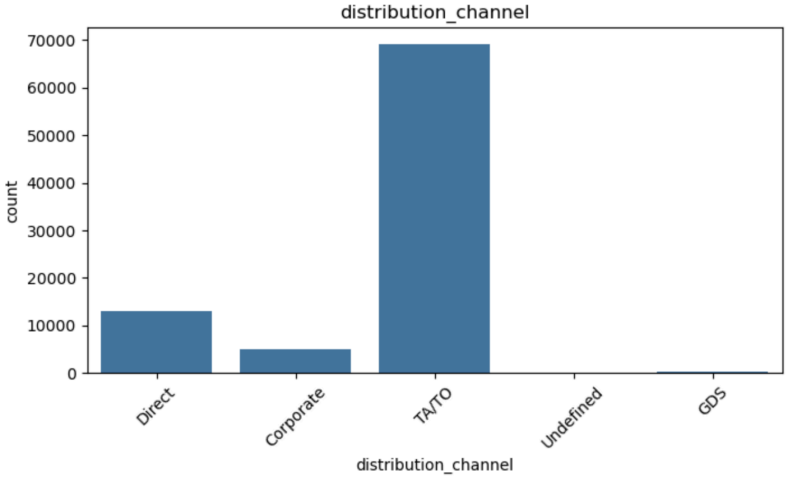
Here we can see that the top 5 correlated features are Lead Time, Market Segment, Distribution Channel, Deposit Type, and ADR. In further investigation of these specific features we can easily see that reservations booked further in advance exhibited a higher cancellation rate, while bookings made closer to the arrival date were less likely to cancel. We can also see that reservations made through online travel agencies showed a higher cancellation rate compared to direct bookings, highlighting the relative value of direct channels. However the visualizations for Distribution Channel, Deposit Type, and ADR yield no visible correlation or pattern. In fact, distribution channel seems to be closely related to Market Segment, and Deposit Type seems to be almost entirely No Deposit. **Based on these observations, all available features were retained for modeling to maximize predictive performance.**

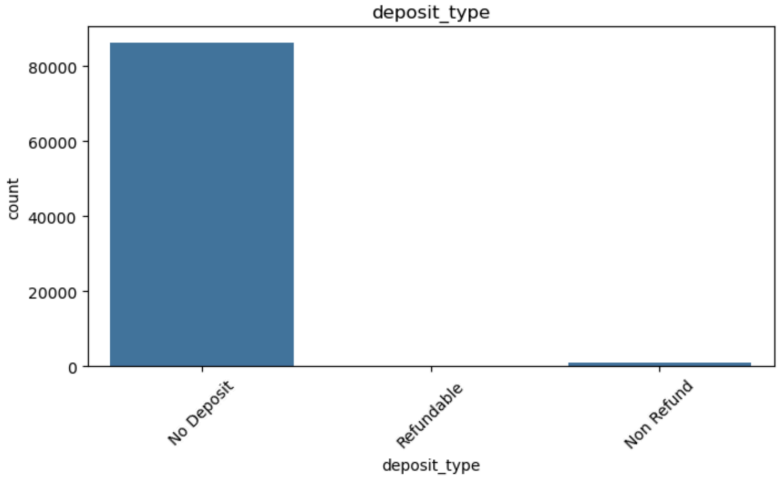












With the insights gathered from the Exploratory Data Analysis phase, the data was ready to be processed for modeling. First, categorical and numerical features were separated. Then categorical variables were converted into dummy variables using

one-hot encoding, while numerical features were scaled using StandardScaler. Lastly, the dataset was split into training and testing sets using train_test_split.

Four classification models were trained and evaluated in the modeling step.

1. Logistic Regression
2. Gradient Boosting Classifier
3. Decision Tree
4. Random Forest

Each model was individually scored, then the scores were pulled together to compare. Model performance was assessed using accuracy, precision, F1-score, and ROC-AUC.

Model	**Accuracy**	**Precision**	**F1**	**ROC-AUC**
Logistic Regression	0.829519	0.764877	0.638788	0.861251
Gradient Boosting Classifier	0.833295	0.766413	0.651185	0.892918
Decision Tree	0.861041	0.761099	0.740353	0.817474
Random Forest	0.900172	0.939655	0.789378	0.966561

The Random Forest model outperformed the other models on all evaluation metrics, making it the clear choice as the final selected model. With a 90% accuracy, we can also be confident that the Random Forest model not only outperformed the others, but will be a reliable source for predicting cancellations.

In conclusion, I recommend that hotel revenue management teams apply this model to new reservations to inform overbooking strategies and mitigate the financial impact of cancellations. By incorporating cancellation probabilities into operational decision-making, hotels can more confidently oversell inventory while reducing the risk of guest displacement. The ultimate objective is to achieve optimal occupancy despite expected cancellations.