Hotel Reservation Booking Cancellation Prediction

Presented by: Pranav Dange and Vabhavi Tickoo

IS 517 : Methods of Data Science
Renee Hendricks

TABLE OF CONTENT

BACKGROUND

Boost in hotel operations

Till 2022, significant raise in hospitality business (~4 trillion)

Anticipating cancellations can help manage hotels

Previous studies predicted cancellations using different algorithms

Performance dependent on variables

Leveraging previous dataset for choosing best variables.

PAST STUDIES AND LIMITATIONS

Studies have attempted to predict cancellations using various algorithms, including decision trees, random forest, and logistic regression.

Limitations of earlier research was the use of small datasets, which may not accurately represent the diverse range of factors that contribute to cancellations.

The limited scope of previous research has resulted in varied outcomes and limited ability to generalize findings across different contexts.

The lack of focus on specific features that have a high correlation with cancellation, leading to less accurate predictions.

PROJECT OBJECTIVE AND DATASET

OBJECTIVE

- The objective of our project is to develop an accurate prediction model for hotel reservation cancellations using our dataset.
- To achieve this, we will perform exploratory data analysis and pre-process the data, including encoding variables with object datatype and dropping redundant variables.
- We will only use variables that have a significant impact on the prediction.
- We will test different models and tune them to improve their accuracy.

DATASET

- Our dataset is the "Hotel Reservations Classification Dataset" from Kaggle, which includes 36275 records of hotel reservations.
- The dataset contains various attributes.
- Both categorical and numerical data are included, which will be further processed according to our data analysis.

PROJECT OBJECTIVE AND DATASET

Booking_ID	Unique identifier of each booking
no_of_adults	Number of adults
no_of_children	Number of children
no_of_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay
no_of_week_nights	Number of week nights (Monday to Friday) the guest stayed or booked to stay
type_of_meal_plan	Type of meal plan booked by the customer
required_car_parking_space	Does the customer require a car parking space? (0 - No, 1 - Yes)
room_type_reserved	Type of room reserved by the customer
lead_time	Number of days between the date of booking and the arrival date
arrival_year	Year of arrival date
arrival_month	Month of arrival date
arrival_date	Date of the month
market_segment_type	Market segment designation
repeated_guest	Is the customer a repeated guest? (0 - No, 1 - Yes)
no_of_previous_cancellations	Number of previous bookings that were canceled by the customer prior to the current
no_of_previous_bookings_not_canceled	Number of previous bookings not canceled by the customer prior to the current
Avg_price_per_room	Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
no_of_special_requests	Total number of special requests made by the customer
booking_status	Flag indicating if the booking was canceled or not

PROPOSED METHOD

Our proposed method for predicting hotel reservation cancellations will consist of several steps, including data preprocessing, feature selection, model selection, and model evaluation.

First pre-process the data

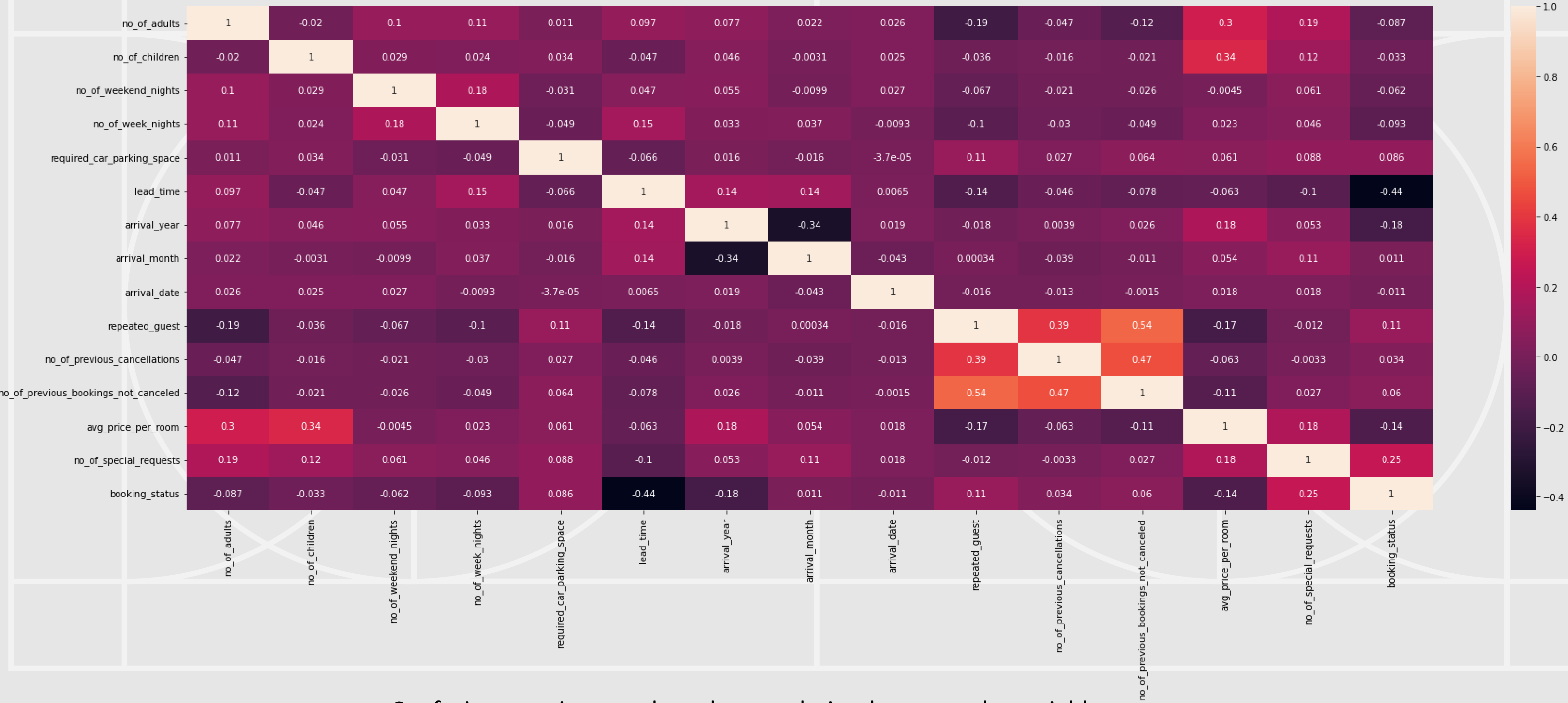
Next, we performed feature selection

Evaluate several machine learning algorithms

After selecting the best algorithm, we will tune its hyperparameters

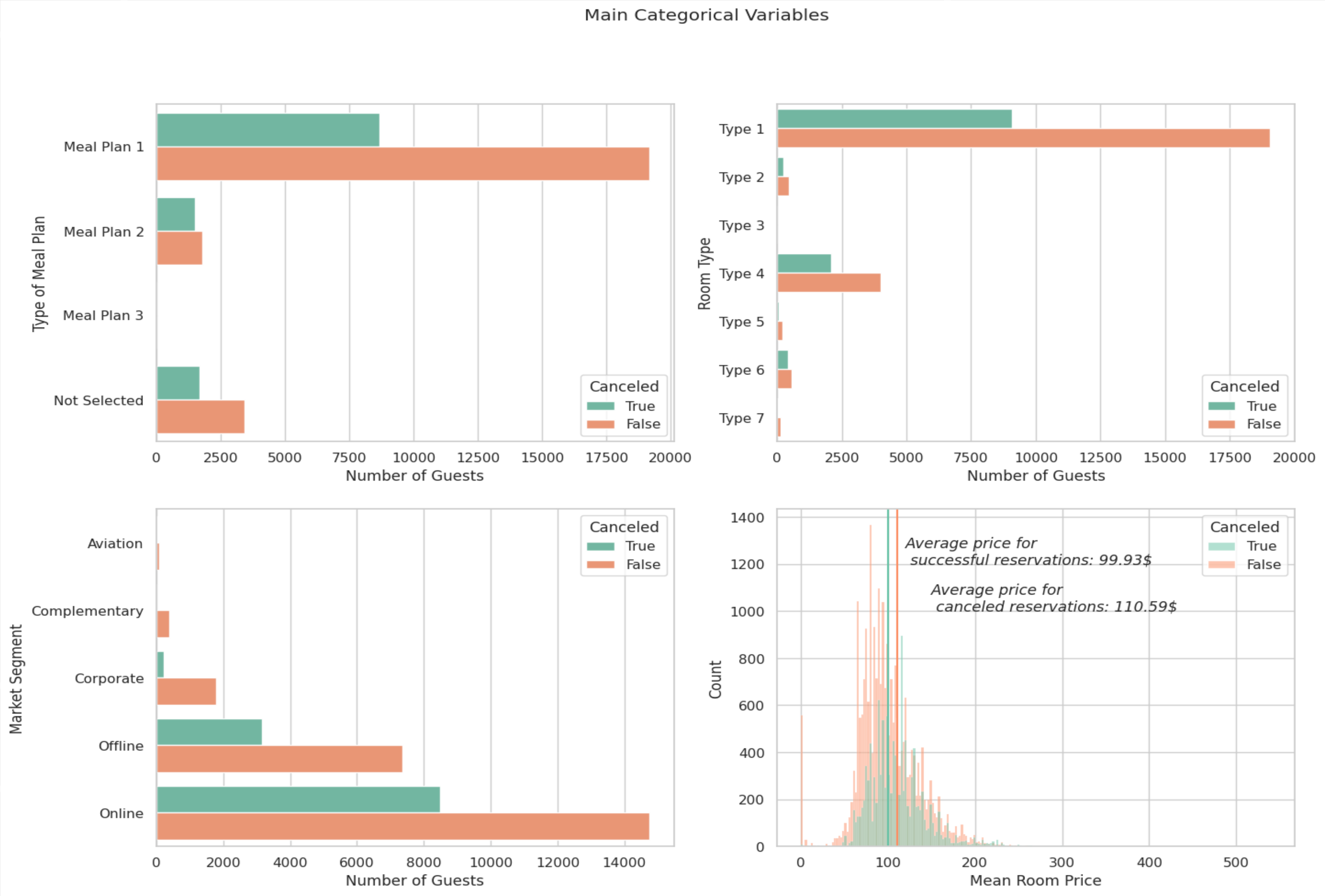
Finally, we will evaluate the performance of our model using various metrics

EXPLORATORY DATA ANALYSIS



Confusion matrix to analyze the correlation between the variables

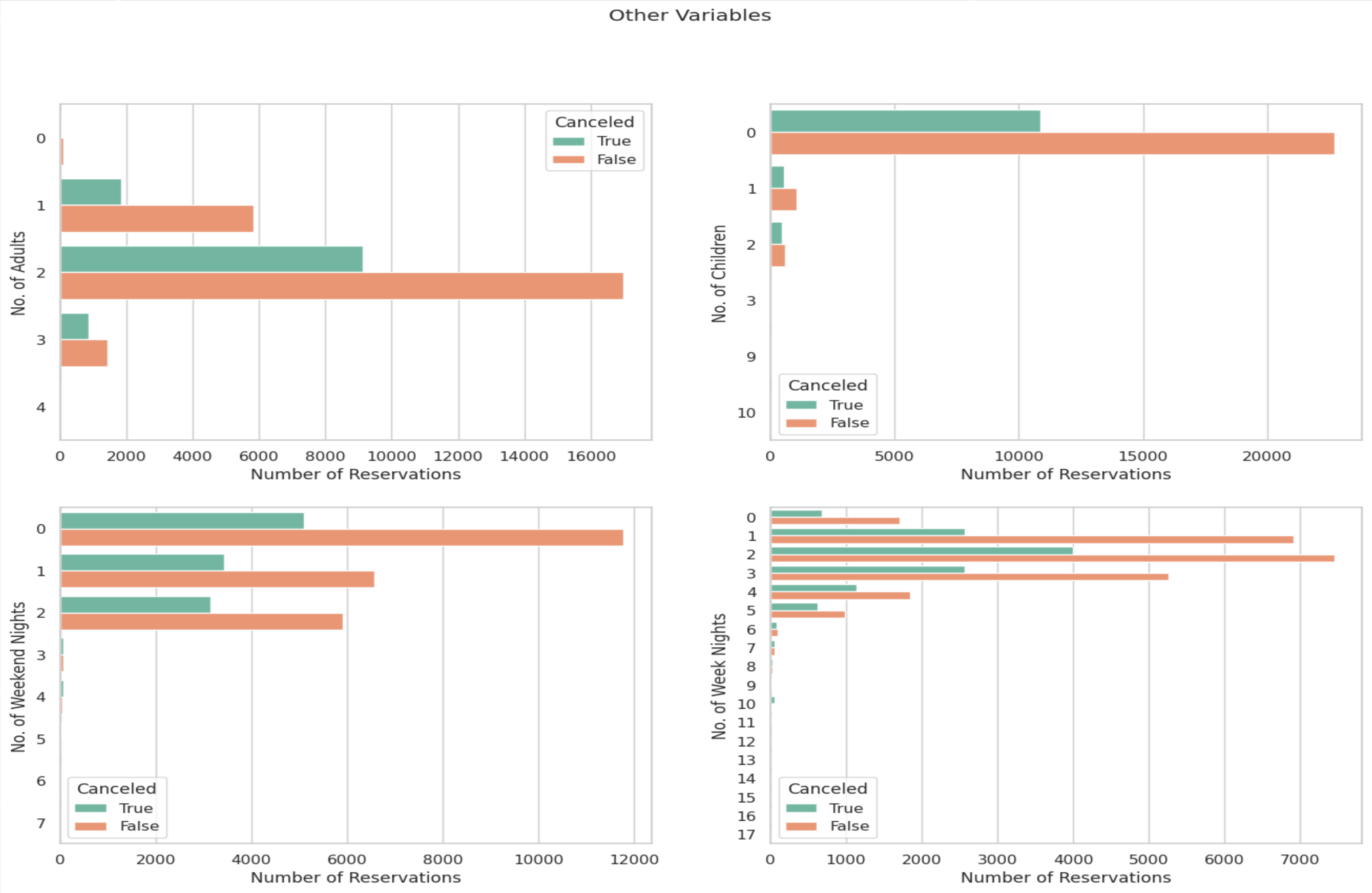
EXPLORATORY DATA ANALYSIS



Upon analysis, we observe that:

- The proportion of canceled bookings remains relatively stable over time.
- Rooms of type 6 have a higher cancellation rate compared to other room types.
- The proportion of offline and corporate cancellations is relatively lower compared to online cancellations.
- Canceled bookings have a higher average price of 110.59 USD compared to successful bookings.

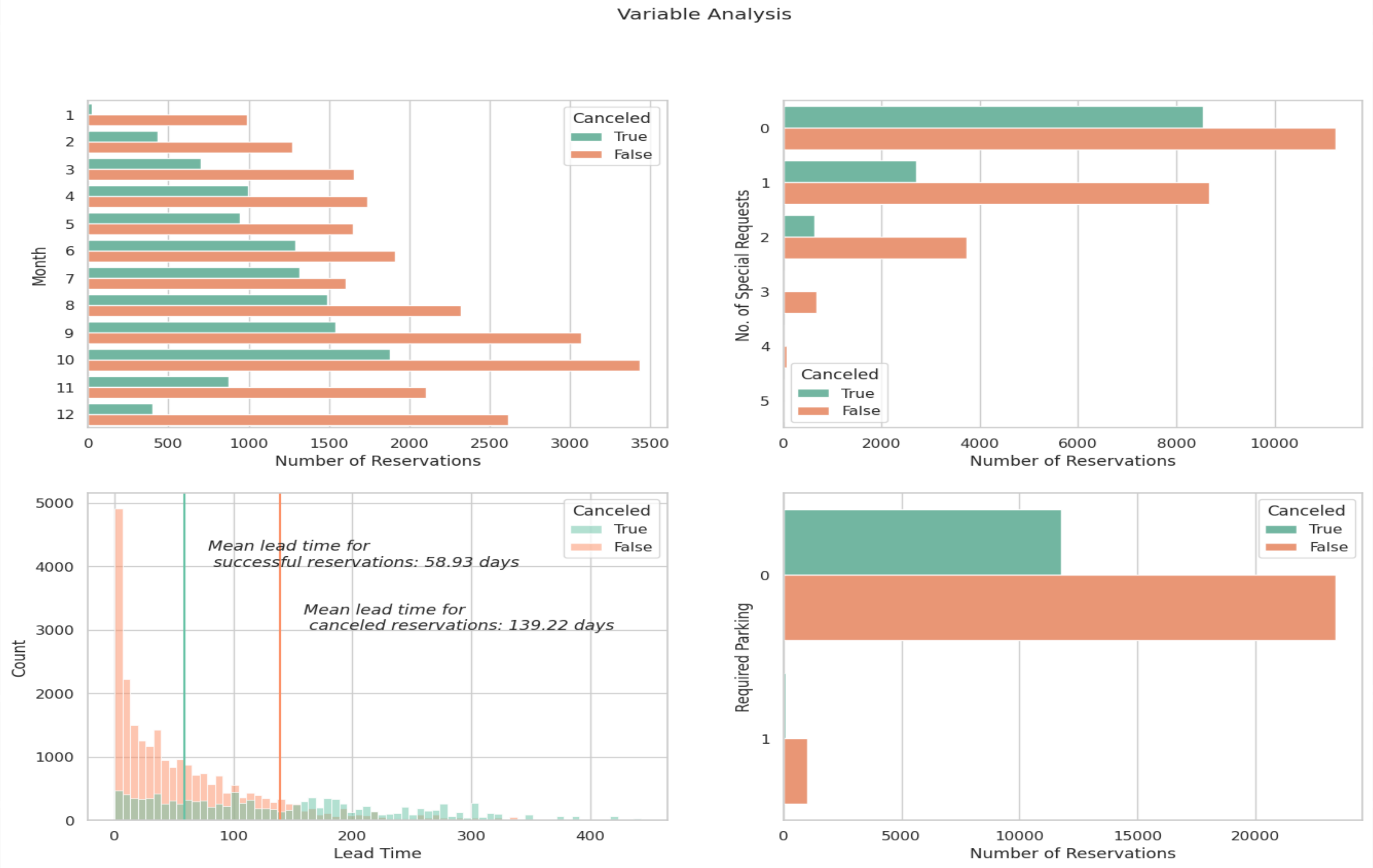
EXPLORATORY DATA ANALYSIS



We also found that:

- The reservations for 3 adults have a higher cancellation rate.
- Similarly, reservations with two children also have a higher likelihood of cancellation.
- Our analysis suggests that there is a positive correlation between the number of weeknights booked and the probability of cancellation.

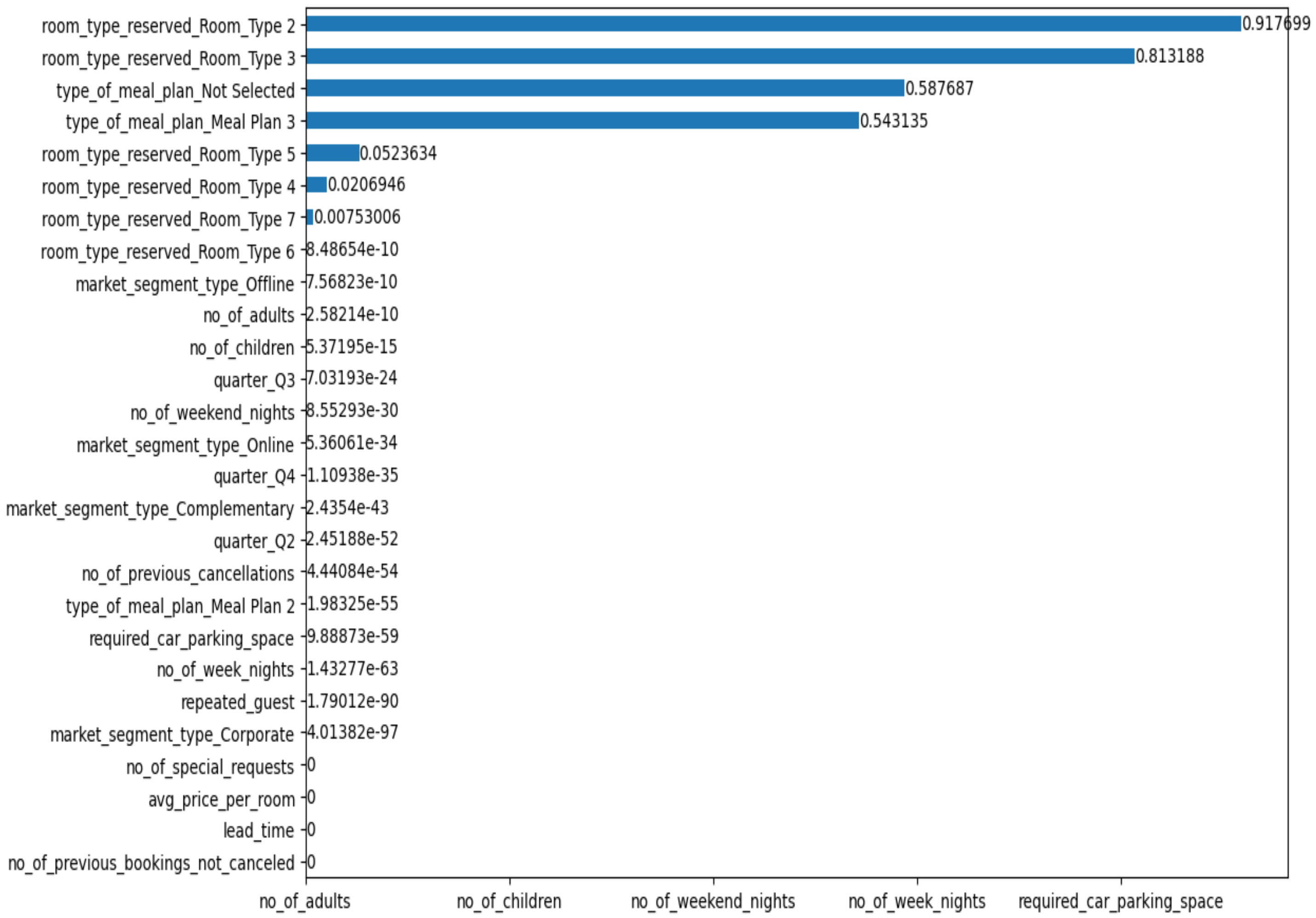
EXPLORATORY DATA ANALYSIS



Observations:

- The proportion of cancellations decreases during winter months.
- There seems to be an inverse relationship between the number of special requests and cancellations.
- Reservations with a longer lead time are more likely to be canceled.
- Reservations that require parking have a very low rate of cancellations.

DATA CLEANING AND FEATURE SELECTION



We pre-processed our data by dropping the unnecessary columns like booking id and arrival year and converted arrival month to quarters and encoding categorical variables.

In feature selection we calculated the chi-squared statistic for each feature to identify the most relevant features. The results were plotted as a horizontal bar chart, with features sorted in ascending order of p-values.

Next, we removed the less important features, such as room types and meal plans that were not selected, to reduce the dimensionality of the dataset. This step was necessary to prevent overfitting and improve model performance.

Features dropped: booking_id, arrival_year, arrival_date, arrival_month, room_type_reserved_Room_Type 2, room_type_reserved_Room_Type 3, type_of_meal_plan_Not Selected, type_of_meal_plan_Meal Plan 3, room_type_reserved_Room_Type 5

MACHINE LEARNING MODELS

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
BaggingClassifier	0.88	0.87	0.87	0.88	1.89
RandomForestClassifier	0.89	0.87	0.87	0.89	3.88
XGBClassifier	0.89	0.86	0.86	0.89	2.69
ExtraTreesClassifier	0.88	0.86	0.86	0.88	2.72
DecisionTreeClassifier	0.86	0.84	0.84	0.86	0.17
LGBMClassifier	0.87	0.84	0.84	0.87	0.53
LabelSpreading	0.83	0.81	0.81	0.83	73.30
LabelPropagation	0.83	0.81	0.81	0.83	31.79
KNeighborsClassifier	0.84	0.81	0.81	0.84	4.09
ExtraTreeClassifier	0.83	0.80	0.80	0.83	0.06
SVC	0.83	0.77	0.77	0.82	38.25
AdaBoostClassifier	0.80	0.75	0.75	0.79	3.08
NearestCentroid	0.75	0.75	0.75	0.76	0.12
NuSVC	0.81	0.74	0.74	0.80	53.20
LogisticRegression	0.79	0.73	0.73	0.78	0.12
CalibratedClassifierCV	0.79	0.73	0.73	0.78	15.00
LinearSVC	0.79	0.72	0.72	0.78	3.06
LinearDiscriminantAnalysis	0.78	0.71	0.71	0.77	0.19
RidgeClassifierCV	0.78	0.71	0.71	0.77	0.10
RidgeClassifier	0.78	0.71	0.71	0.77	0.07
BernoulliNB	0.76	0.71	0.71	0.76	0.23
SGDClassifier	0.78	0.70	0.70	0.77	0.18
PassiveAggressiveClassifier	0.65	0.68	0.68	0.67	0.08
Perceptron	0.66	0.64	0.64	0.67	0.08
QuadraticDiscriminantAnalysis	0.45	0.59	0.59	0.40	0.10
GaussianNB	0.44	0.58	0.58	0.38	0.05

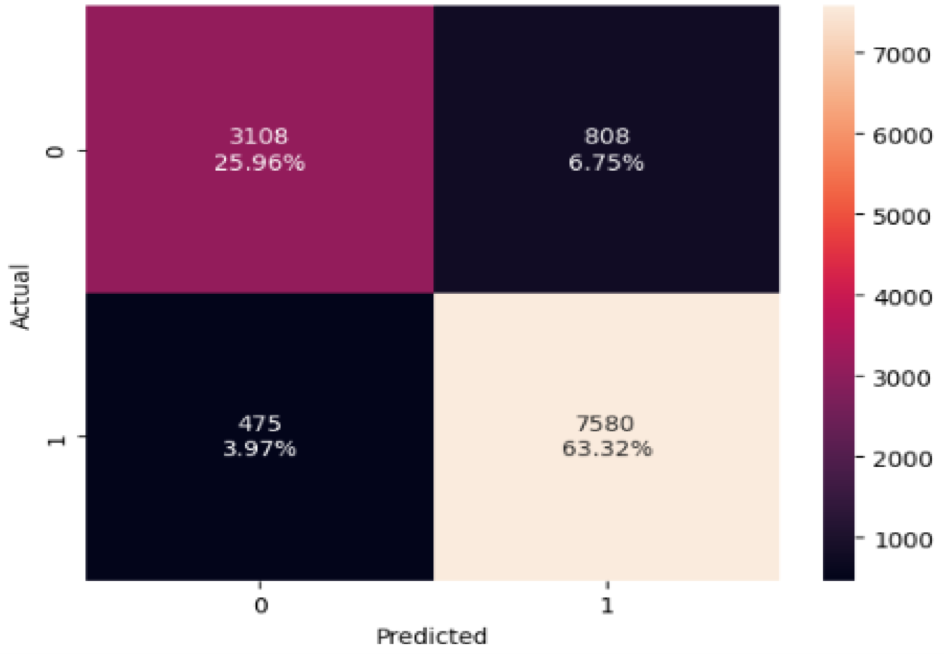
Initially we just encoded the categorical features and did an 80:20 split and used lazy predict to get an overview of how the different models would perform.

Out of the 26 models we analysed the best performing methods were Random Forest Classifier and XGBClassifier which had accuracy of around 89%and Balanced Accuracy of around 87%.

So, our approach was to fine tune these models.

MACHINE LEARNING MODELS

- We performed different data pre-processing techniques like dropping the columns with low correlation on the result variable and even performed over sampling on the data because there was a skew in the cancellation and non-cancellation value. However, we were not able to improve on the 89% Accuracy of the model.



	Model	Accuracy Score	F1 score	Precision	Recall
0	XGBoost	0.892824	0.921973	0.903672	0.94103

MACHINE LEARNING MODELS

First columns names are converted to lowercase, and unnecessary columns such as booking ID, arrival year, and arrival date are dropped because it did not have impact on the predicted value based on the confusion matrix we analysed in our EDA.

We created a new column named 'quarter' using the arrival month column values, and the arrival month column is dropped.

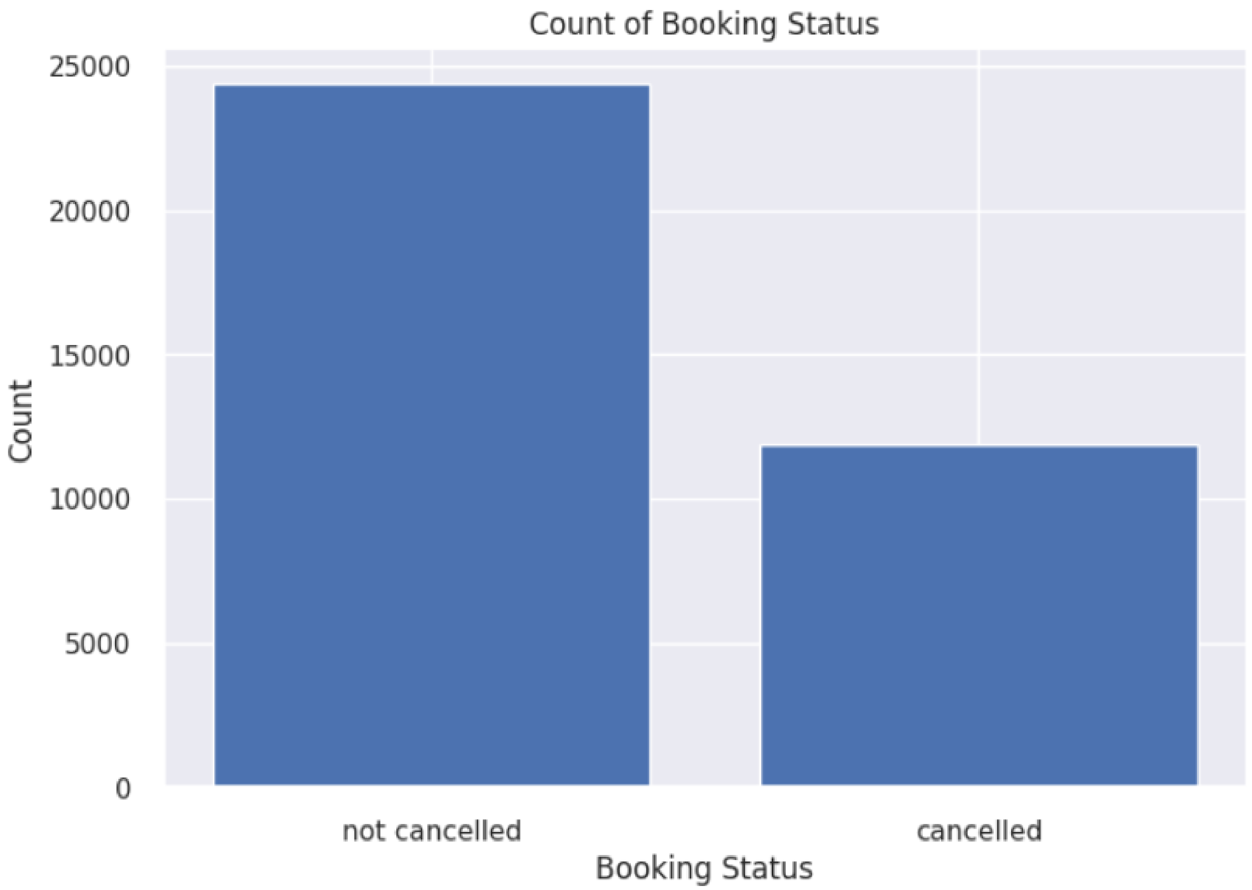
We encoded categorical variables using one-hot encoding, and the target variable is separated from the feature set.

We performed Chi-square test to select the most significant features for the classification task.

Features with high p-values are dropped.

As our count was not balanced for the cancelled and not cancelled data, we performed SMOTE oversampling for our Random Forest Classifier.

MACHINE LEARNING MODELS



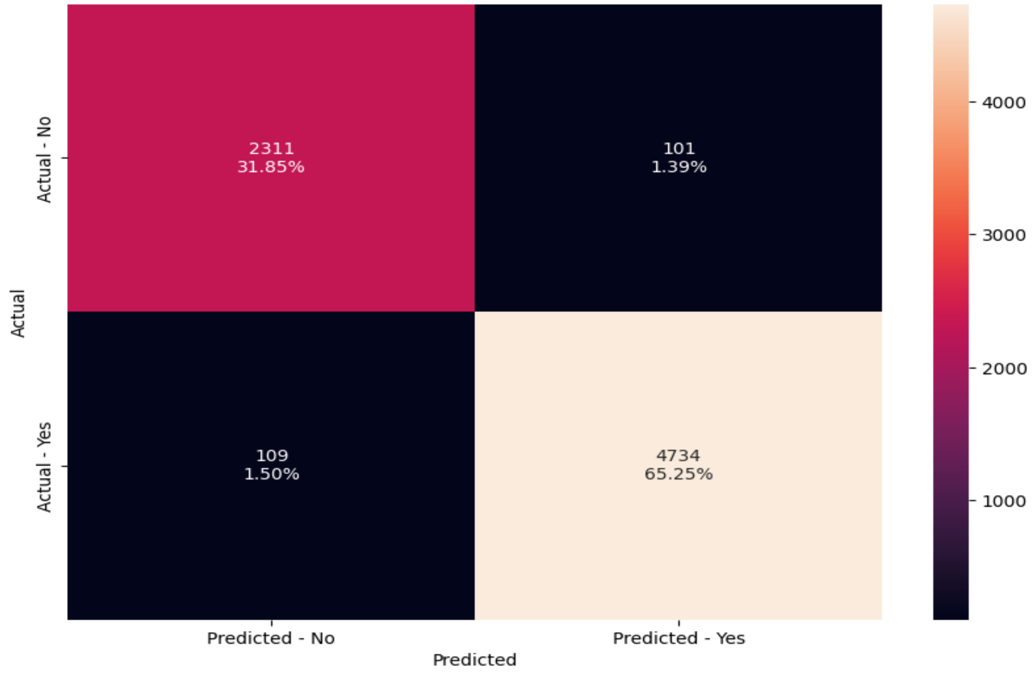
- We used SMOTE to address the class imbalance problem in the dataset.

```
df['booking_status'].value_counts()
```

```
1    24390
0    11885
Name: booking_status, dtype: int64
```

MACHINE LEARNING MODELS

Result of our Random Forest Classifier.

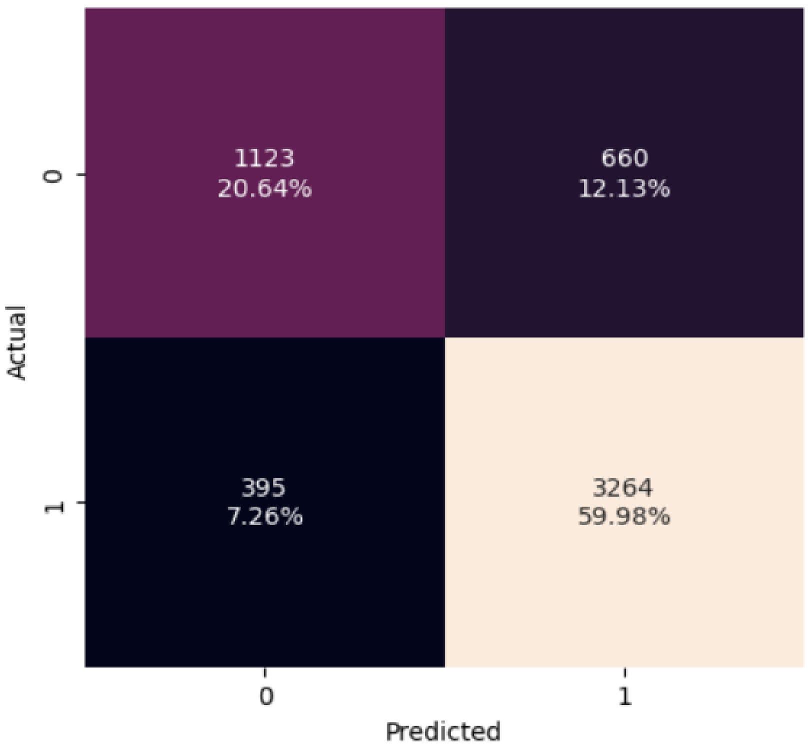



```
Accuracy on training set :  0.9735010337698139
Accuracy on test set :  0.9710544452101999
Recall on training set :  0.9797922954929145
Recall on test set :  0.977493289283502
Precision on training set :  0.9808460514186214
Precision on test set :  0.979110651499483
```

	Model	Accuracy Score	F1 score	Precision	Recall
0	Random Forest Classifier	0.971054	0.978301	0.979111	0.977493

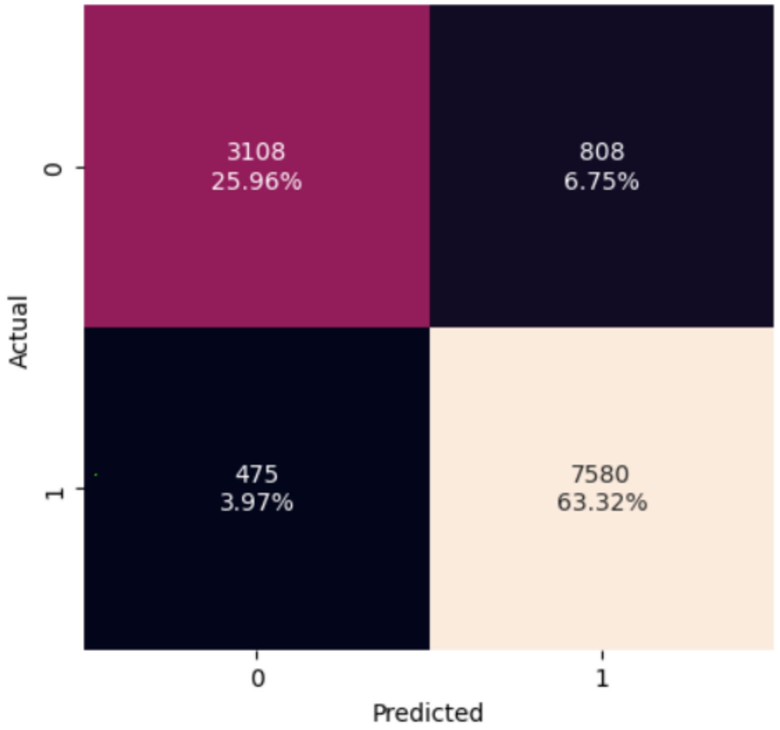
RESULT COMPARISONS

Logistic Regression



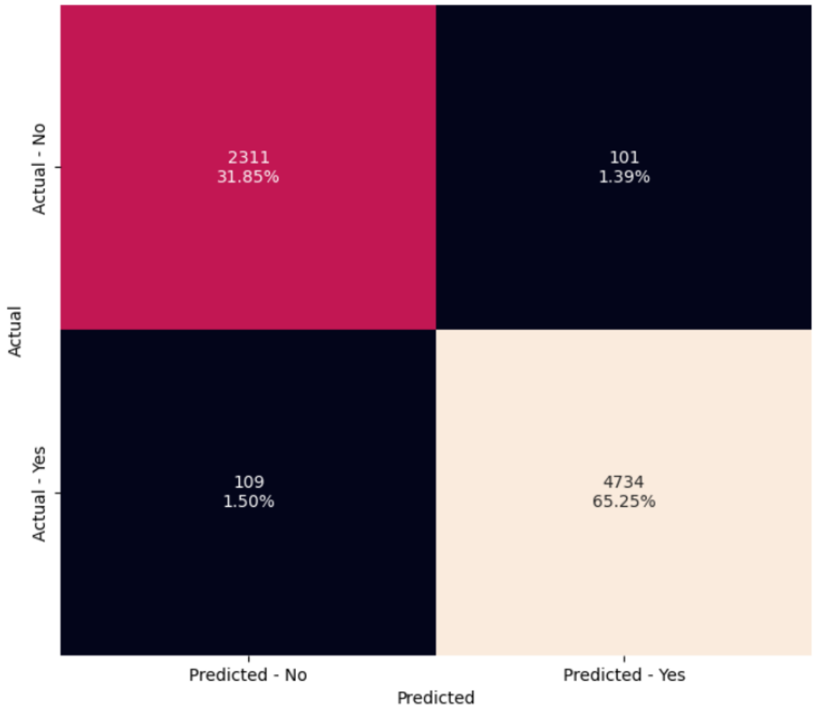
	Model	Accuracy Score	F1 score	Precision	Recall
0	Logistic Regression	0.806137	0.860873	0.831804	0.892047

XGBoost



	Model	Accuracy Score	F1 score	Precision	Recall
0	XGBoost	0.892824	0.921973	0.903672	0.94103

Random Forest



	Model	Accuracy Score	F1 score	Precision	Recall
0	Random Forest Classifier	0.971054	0.978301	0.979111	0.977493

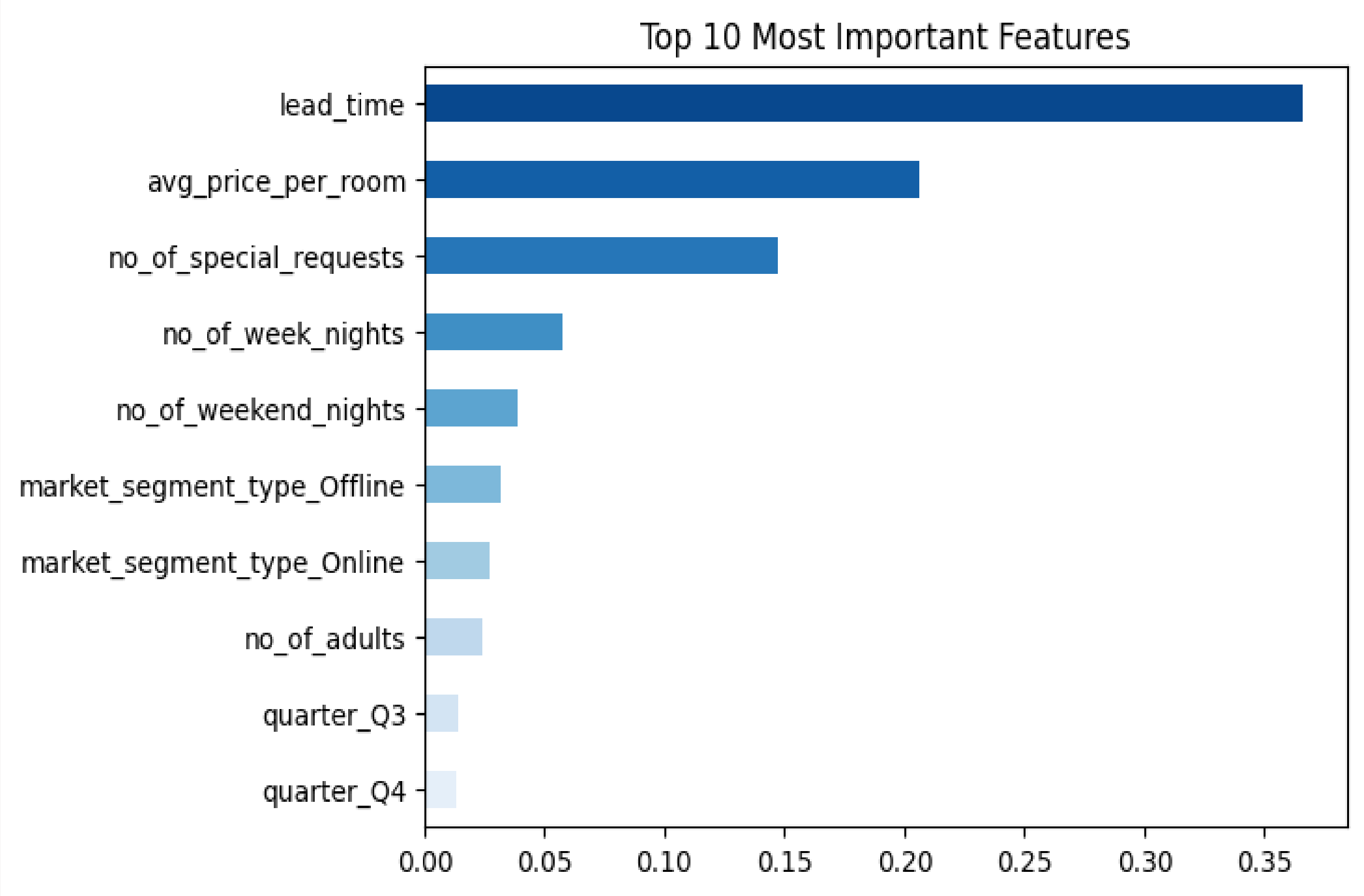
RESULT COMPARISONS

Model	Accuracy	F1 Score	Precision	Recall
Logistic Regression	80.49	85.96	83.29	88.81
Naive Bayes	67	76	51	61
SVM Classifier	76.01	76	75	78
Decision Tree Classifier	84	86	84	89
XGBoost	89.29	92.19	90.37	94.10
Random Forest	**97.10**	**97.83**	**97.91**	**97.75**

LIMITATIONS

- Factors outside our control could impact prediction accuracy
- Limited dataset in terms of time period and region
- Dataset may contain errors or inconsistencies
- Model may not be generalizable to other hotel properties
- Limited by quality and quantity of data available
- Additional variables may impact cancellations that are not captured in our dataset
- Limitations in algorithm performance and feature selection.

CONCLUSION



1. Lead_Time, Avg_price_room, and Number of special requests are the top 3 variables for predicting cancellations.
2. The Random Forest model performs the best with a 97.10% accuracy rate.
3. The number of nights is also a significant factor, but the model assigns less importance to it.
4. Market segments also impact cancellation probability. This is related to how the booking is done.

References

- Dataset: https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset
- https://ieeexplore-ieee-org.proxy2.library.illinois.edu/stamp/stamp.jsp?tp=&arnumber=9299011
- https://reader.elsevier.com/reader/sd/pii/S2352340918315191?token=151498555C6714B58A822BC9AECD1115C6CCD0604F26D5390A6BE10D712D9CB2C8281231F7348830DBF401C5FD0513F6&originRegion=us-east-1&originCreation=20230323003607
- https://ia-institute.com/wp-content/uploads/2021/07/IAI-Journal_2.2021.pdf
- https://www.kaggle.com/code/raphaelmarconato/hotel-reservations-eda-balancing-and-ml-93-4

QUESTIONS?