Project Proposal

Data-Driven Forecasting of Hotel Booking Cancellations

Pranav Dange

Vabhavi Tickoo

1) Background/History

1.1. Literature Review of Dataset or Analysis Method [past research question(s)]

By 2023, there will be at least 187,000 hotels operating globally. Worldwide, there are about 17.5 million guest rooms. As of 2022, the worldwide hospitality market is estimated to be valued at about $4.548 trillion. The ability to anticipate cancellations may help hotels manage their inventory, streamline operations, improve the guest experience, and improve profitability. There are a lot of reasons that could impact the cancellation; however, we would try to predict the cancellation based on the variables present in our dataset. There have been several studies done that have tried to predict cancellation using different algorithms, such as random forest, decision trees, and logistic regression. However, the performance of these algorithms depends on the features present in the dataset. In our project, we will leverage the insights of the previous research and try to see how we could use this knowledge to choose the best features from our dataset to predict the cancellation with high accuracy.

1.2. Limitations of previous studies or analysis methods

The previous study had different variables that were used to predict their results; however, our dataset has different values that we would use to predict the cancellation. The use of tiny datasets was one of the limitations of earlier research. Thus, to create more precise prediction models, more and more variables with significant correlations to the prediction values are required. In order to obtain a much more precise prediction on reservation cancellation, all essential elements will be investigated, and the factors having the highest significance will be fed to the model to improve the prediction.

2) Proposed Project

2.1. Objective

The main objective of this proposed project is to develop an accurate prediction model for hotel reservation cancellations for our dataset. We will perform exploratory data analysis and perform the necessary data preprocessing steps. Also, we will need to encode the variables that have a datatype as an object before we feed them to the model, use only the variables that have a significant impact on the prediction, and drop the redundant variables. After that, we will try different models, test their accuracy, and tune the model to improve its accuracy.

2.2. Dataset(s)

The dataset used for this project is the "Hotel Reservations Classification Dataset" available on Kaggle, containing 119,390 records of hotel reservations made by customers, including various attributes such as reservation status, lead time, arrival date, customer demographics, and the cost of the room. The data contains categorical and numerical data that will be further processed according to our data analysis.

2.3. Proposed Method(s) Applied

To properly anticipate hotel reservation cancellations, we will test different algorithms on our data. Among the approaches offered are:

Decision Trees: Decision trees are a common approach for classification. They operate by recursively dividing the data into smaller groups based on the value of a specific property. This technique can assist in determining the most important variables influencing cancellations in our dataset.

Random Forest: This approach uses numerous decision trees to increase prediction accuracy. It avoids overfitting and can be used to predict cancellation in our case. We will evaluate this approach as well for our data.

Logistic Regression: Logistic regression is a popular approach for binary classification issues that can give insights into the factors that influence cancellations and give us if a particular booking will be canceled or not.

XGBoost: This approach combines numerous weak models to generate a powerful prediction model. It operates by creating decision trees repeatedly, with each new tree correcting the faults of the prior one. This approach can manage missing values, and feature selection, and can increase model performance when compared to other algorithms. We will evaluate this model as well for our data.

2.4. Evaluation Metric(s)

The evaluation metric used to assess the model's performance will be the Accuracy percentage, AUC (Area Under Curve), and F1 score and we will also evaluate the number of correct and incorrect predictions in our data. We will choose the model which will best suit our data using the above factors

2.5. Expectation(s) of Results

Based on the research on the previous papers written in the same domain. Which successfully predicted the result with an accuracy of around 80% we expect our approach to provide around this number. However, our data is different and there are a few categorical attributes that we need to encode and feed to our model which could have an impact on our score. Also, there are other factors that could impact our accuracy which we would only encounter once we train our model.

Project Implementation

3.1 Data Cleaning Steps Performed and Outcome

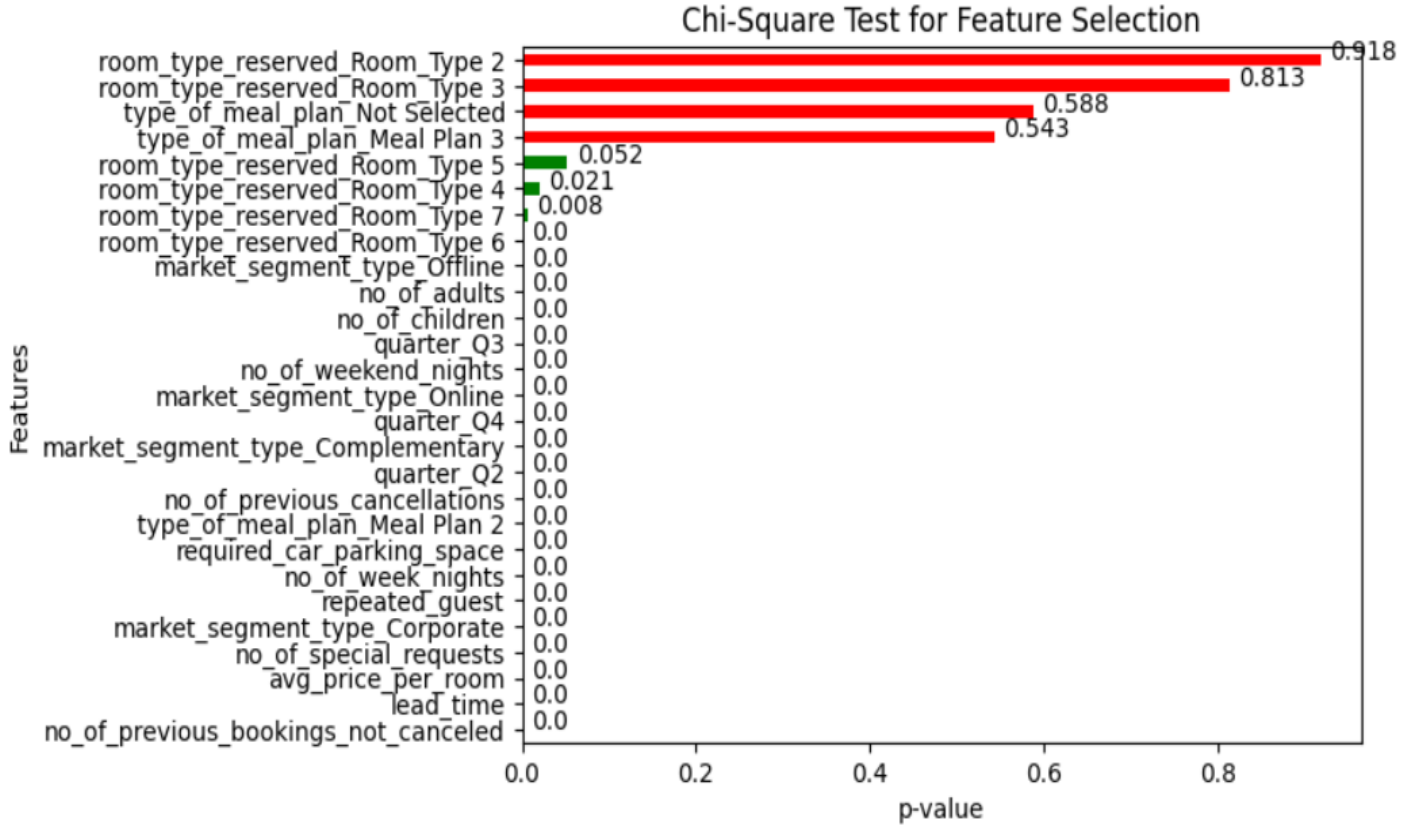
In terms of data cleaning, we were fortunate that there were no null values in our dataset. Hence, we were not required to handle this scenario. However, we had prepared that we would replace the numerical data with the median values in case we would have encountered this scenario. In data cleaning the first step we performed was to convert all the column names to lowercase. Then we found out that booking_id, arrival_year, and arrival_date had the least amount of correlation on the predicted value. We found this correlation using the

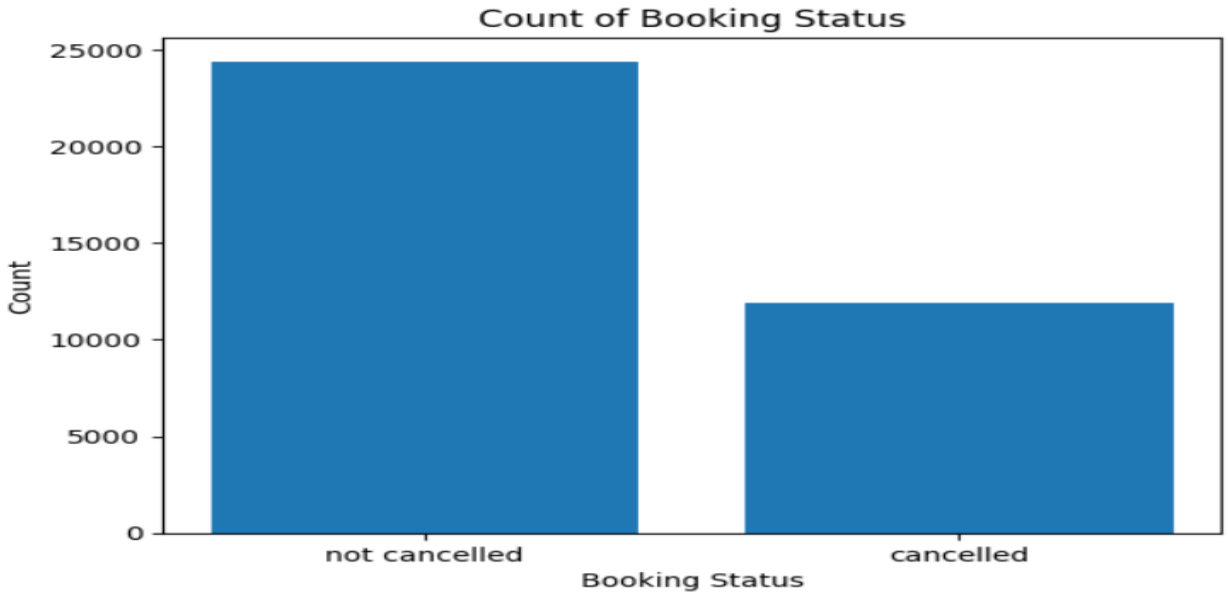
confusion matrix. We also encountered a surge in hotel reservations in certain months in our EDA and this surge was experienced in certain quarters. One of the reasons for this surge would be the hotel season in the December. So we split the arrival_month column into four quarters and added a new column feature called quarter to our data. We had certain categorical features in our dataset; however, we needed to encode this before feeding it to our model as this was the only way our model could use this feature. So to handle this we performed one hot encoding on our categorical features.

One of the drawbacks of the previous approach was that all the previous research had not focused on feeding the most important features to the model. So our approach to solving this problem was a chi-square test to plot the most important features in ascending order of p values and drop those features where p values were higher than 0.5. This way we fed our model with only the most relevant features.

One more thing which we encountered was that the prediction value count was not equally distributed in our dataset. So to handle this scenario we used the SMOTE over-sampling technique.

The outcome of all the steps performed above was that we only feed our model with the most relevant features and the data was balanced which would help the model to understand the features in a better way while training.






```
Not_Canceled    24390
Canceled        11885
```

3.2 Exploratory Data Analysis:

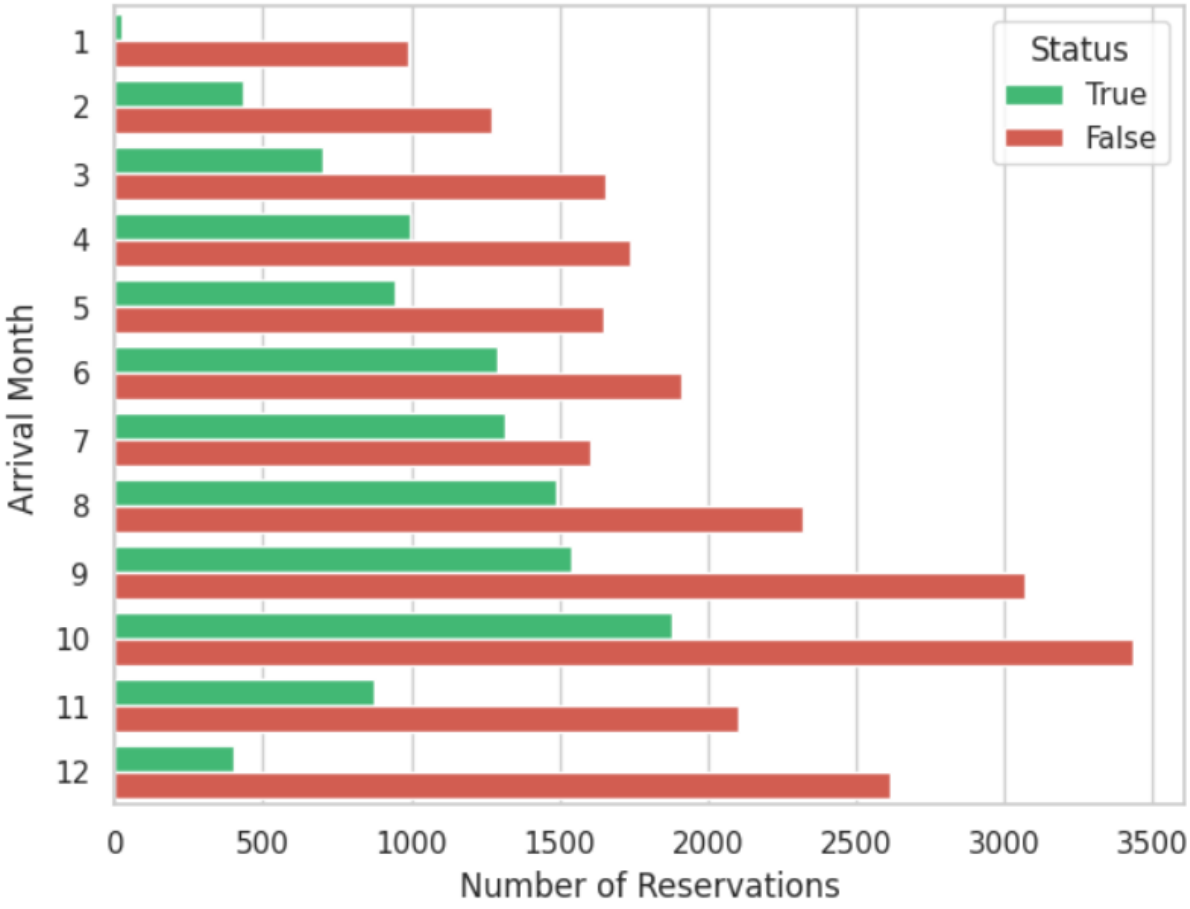
In the EDA we analyzed the correlation between all the features using the confusion matrix and dropped the features which had the least values in the data-cleaning steps. Furthermore, we also analyzed the arrival patterns in terms of the month as we used this to add a new feature called quarter in our data preprocessing step.

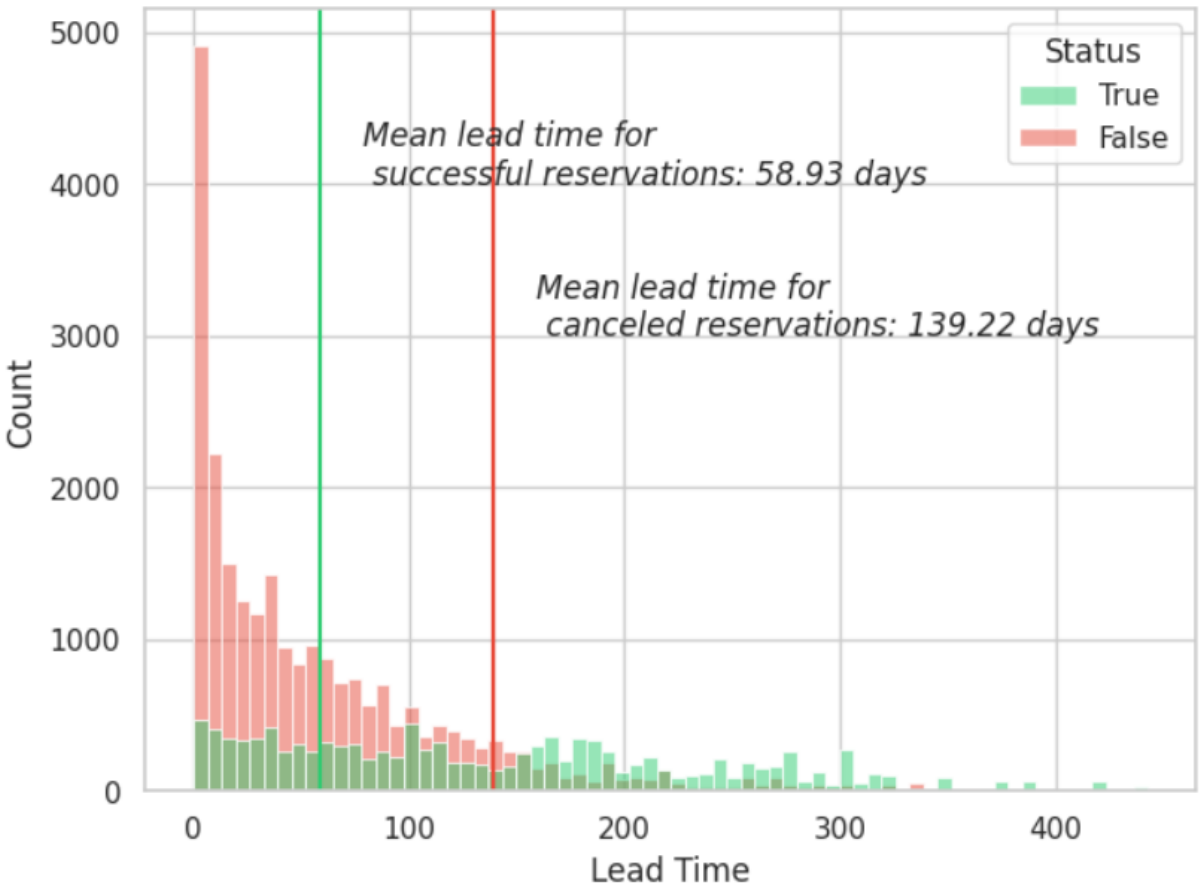
We plotted all the categorical, numerical, and variable features to find any pattern we could use in order to enhance our model. However, the most important discovery in this process was to find the pattern in the lead_time and avg_price of the hotel based on canceled and not canceled values and this plot clearly showed the difference in the average price and lead time between the canceled and not canceled values. The plot clearly shows that the Average Price of successful reservations was lower than its significant other. Moreover, the lead time of successful reservations was significantly lower than that of canceled ones.

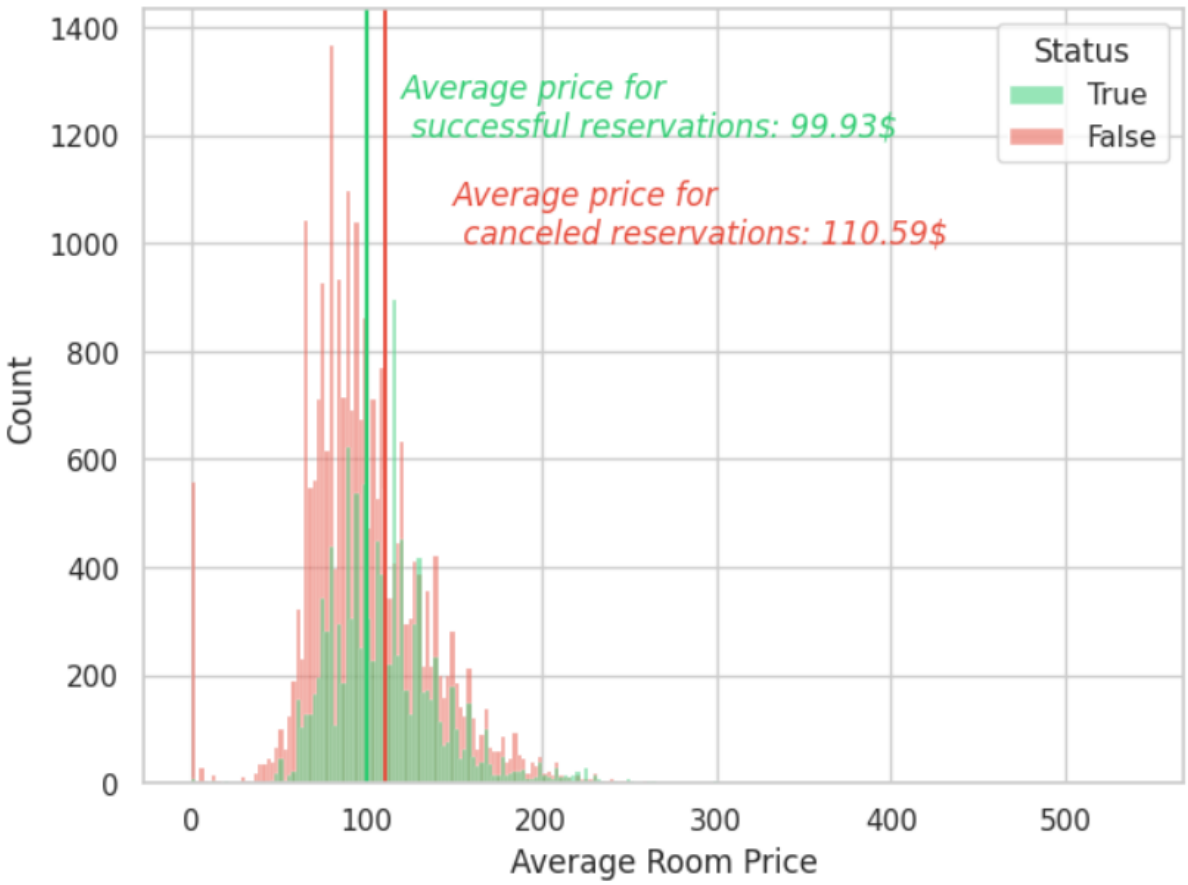
Our EDA also helped us in analyzing certain patterns in our data which were as follows:

1. The proportion of canceled bookings remains relatively stable over time.
2. Rooms of type 6 have a higher cancellation rate compared to other room types.
3. The proportion of offline and corporate cancellations is relatively lower compared to online cancellations.
4. Canceled bookings have a higher average price of 110.59 USD compared to successful bookings.
5. The reservations for 3 adults have a higher cancellation rate.
6. Similarly, reservations with two children also have a higher likelihood of cancellation.
7. Our analysis suggests that there is a positive correlation between the number of weeknights booked and the probability of cancellation.

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



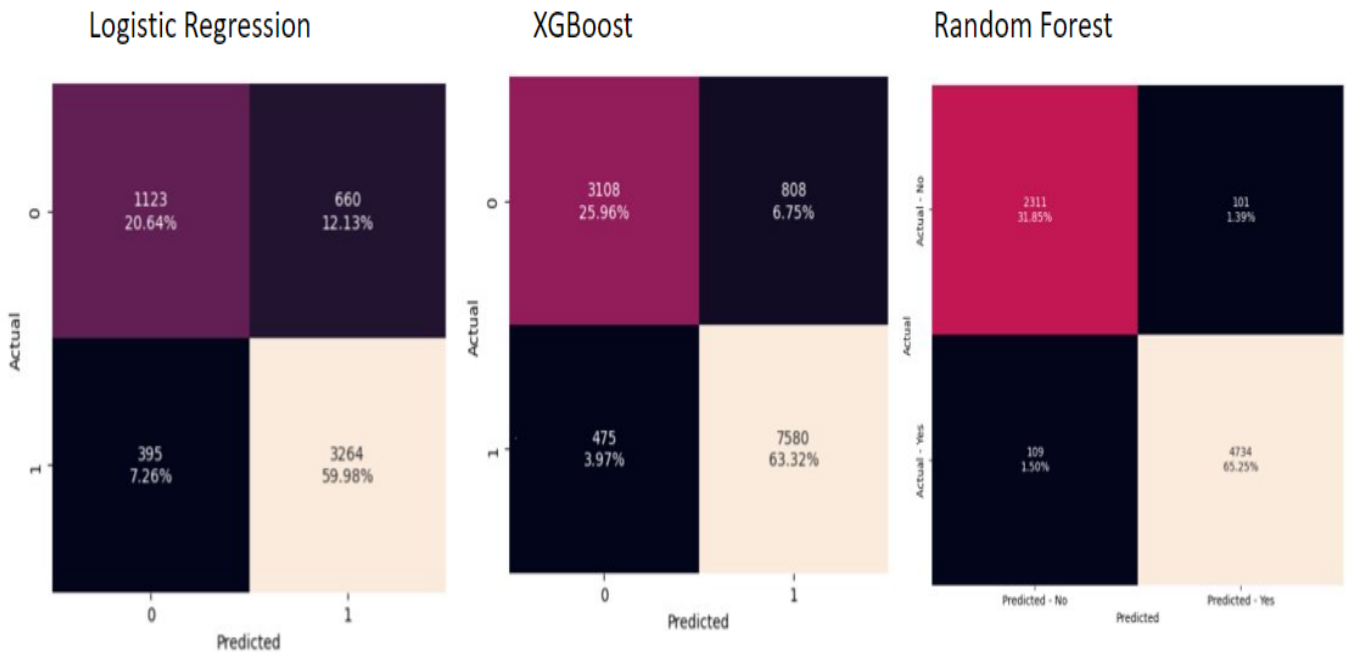

Status
True
False
Mean lead time for
successful reservations: 58.93 days
Mean lead time for
canceled reservations: 139.22 days
Count
Lead Time
0
1000
2000
3000
4000
5000
0
100
200
300
400



3.3 Machine Learning Implementations and Results

Once we had cleaned our data and selected only the relevant features to feed the model, we split our training and testing data in an 80:20 split and used a Python library called Lazy Predict to analyze the 26 classification algorithms. We found that the Random Forest and the XGBoost Classifier performed the best for our model. Hence, we decided to go ahead with random forest classification. To optimize this algorithm, we calculated the best values to feed our model using GridSearch, which calculated the best n_estimator value as 155 and fed this to our model.

In terms of results, we can see that our model is significantly better than the other approaches that were previously applied.



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**

3.4 Limitations

We were able to improve the accuracy in comparison to the previous research. However, there are certain limitation that needs to be addressed which are listed below:

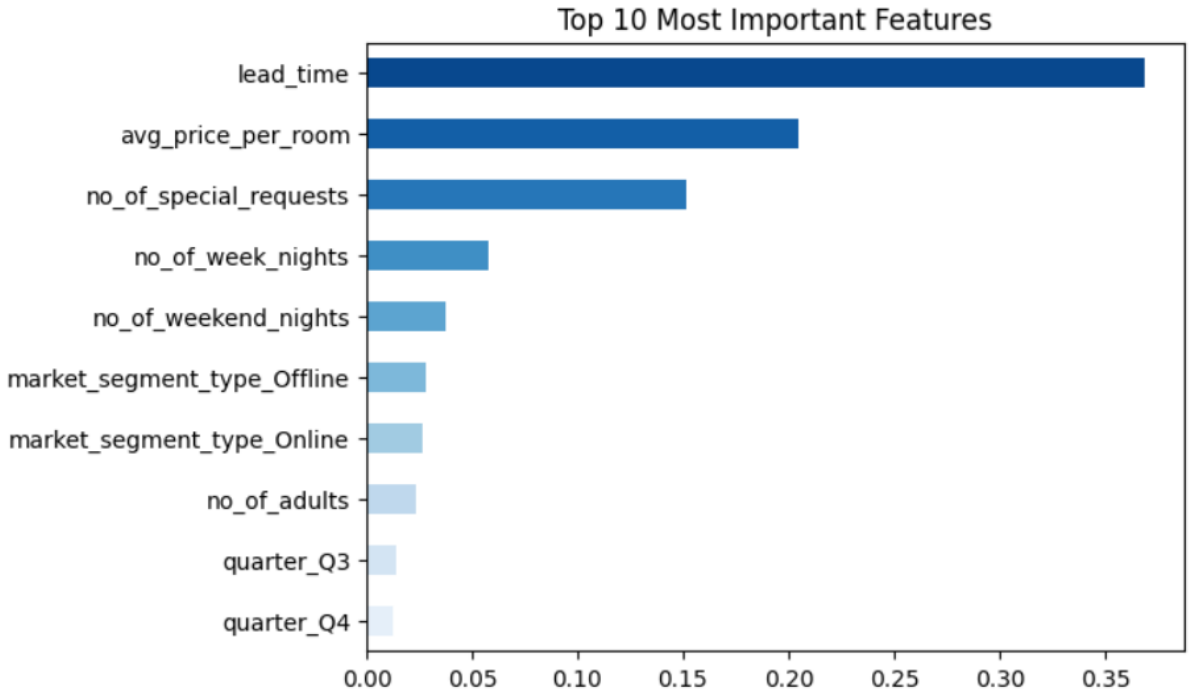
1. There might be certain factors outside the scope of this dataset or something that was not taken into consideration in this dataset that could have an impact on the prediction accuracy.
2. This dataset is restricted to data only from a certain time frame and demographical region hence, we cannot consider this success universal. This is something that needs to be further investigated in order to validate the accuracy of our model further.

3. The dataset which we used cannot be considered error-free. If the amount of error is significant then this could invalidate our results.
4. However, this dataset contained a good amount of rows and we performed oversampling, this cannot be considered a significantly large amount of data and the results which we generated cannot be generalized.
5. There could be certain features that were not considered in this data collection which could have an impact on the result and this could hamper the accuracy of our results.

3.5 Conclusions

We can conclude the following based on our observations:

1. Lead_Time, Avg_price_room, and Number of special requests are the top 3 variables for predicting cancellations. We can conclude this based on the top 10 most Important Features Plot.
2. The Random Forest model performs the best with a 97.10% accuracy rate on our dataset.
3. The number of nights is also a significant factor, but the model assigns less importance to it compared to the other features.
4. Market segments also impact cancellation probability. This is related to how the booking is done.
5. So based on the top most important features we can conclude that based on our EDA the average cost of the room and the Lead Time i.e. the Time between the booking and the actual date of arrival are two of the most significant features for cancellation prediction.



References:

1. https://ieeexplore-ieee-org.proxy2.library.illinois.edu/stamp/stamp.jsp?tp=&arnumber=9299011
2. https://reader.elsevier.com/reader/sd/pii/S2352340918315191?token=151498555C6714B58A822BC9AECD1115C6CCD0604F26D5390A6BE10D712D9CB2C8281231F7348830DBF401C5FD0513F6&originRegion=us-east-1&originCreation=20230323003607
3. https://ia-institute.com/wp-content/uploads/2021/07/IAI-Journal_2.2021.pdf
4. https://www.kaggle.com/code/christophertimmons/random-forest-97-accuracy-score
5. https://www.kaggle.com/code/leandrocassius/predicting-cancellations-with-xgb#Model-Features-&-Conclusion
6. https://www.kaggle.com/code/raphaelmarconato/hotel-reservations-eda-balancing-and-ml-93-4
7. Dataset: https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset