Project on

“Prediction of booking status and Revenue Optimization”

By

Devanshi Mehta

Table of Contents

Introduction 3

Objectives: 4

Data Pipeline 4

Data Collection 4

Data Cleaning 4

Exploratory Data Analysis 5

Data Modeling 11

K-NN 12

Random Forest 13

Discriminant Analysis 13

Model Results 14

Evaluating Model Results 15

Confusion Matrix 15

Discriminant Analysis 15

Conclusion 18

References 19

Introduction

One of the leading industries until COVID used to be the Hotel industry. Whether it is going on vacations, stay-cation or business trips, getting a booking in hotels would be difficult during the peak seasons. Due to the rush or certain other circumstances, reservations would often be cancelled which can potentially cost the hotel another customer as well. In order to keep a track of the business, manage the activities such as reservation status and how cancellations can affect the bookings; this project will aim to provide some insights from the data, implement a predictive model to predict the reservation status based on different factors, and optimize revenue using the predictions.

The main objective is to predict the cancellation of customers booking a room at hotel, and optimizing the revenue of the hotel. The cancellation policy set was that if any customer will cancel booking any time 0-3 days prior of checking in, they have to pay a certain percentage of booking amount as penalty. For prediction of the reservation status or cancellation status, I have implemented two algorithms: K-NN and Random forest classifier. Results from both models have been compared and model with more accurate predictions have been selected for further analysis. The models were implemented using Python notebook and optimization using Discriminant Analysis was done using Excel. Discriminant Analysis used the confusion matrix from the results of the prediction model. It used the matrix to determine whether we incurred profit or any cost has been incurred. If the prediction is accurate, we can determine the reason for cancellation and work on saving the reservation from getting cancelled by offering some extra service or discount. All the visualizations in project have been done using the “matplotlib” library.

Objectives:

1. Predicting the cancellation status of bookings in hotel using two models: KNN, Random Forest
2. Optimize the revenue by reducing the loss/cost and increasing profits

Data Pipeline

Data Collection

Before leaving for any information related excursion, you need to procure information to examine first. Before collecting data, following questions should be kept in mind:

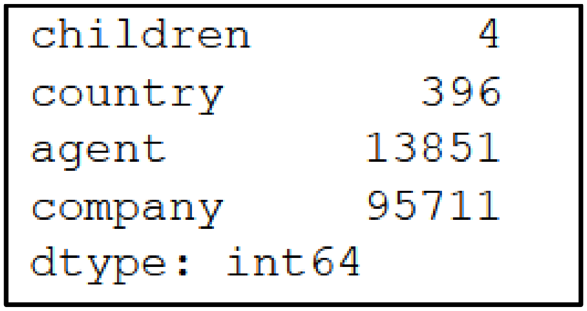
- Which source should I gather the information from?
- What sort of information do I require for the examination that I am going to begin?
- What sort of assortment techniques or channels are accessible to me?

For our project, data was sourced from Kaggle. The timeline in data was that of 3 years. It consisted of 32 columns. All the columns were used for the analysis.

Data Cleaning

This is the way towards recognizing and revising (or eliminating) bad or wrong records from a record set, table, or data set and alludes to distinguishing fragmented, mistaken, off base or immaterial pieces of the information and afterward supplanting, adjusting, or erasing the grimy or coarse information.

The data had four columns which consisted of null values: “Children”, “Country”, “Age”, and “Company”. The null values in these columns were replaced with either 0 (for integer columns) or “no country” for the country column and “no company” for company columns which were string type.



```
children        4
country       396
agent       13851
company     95711
dtype: int64
```

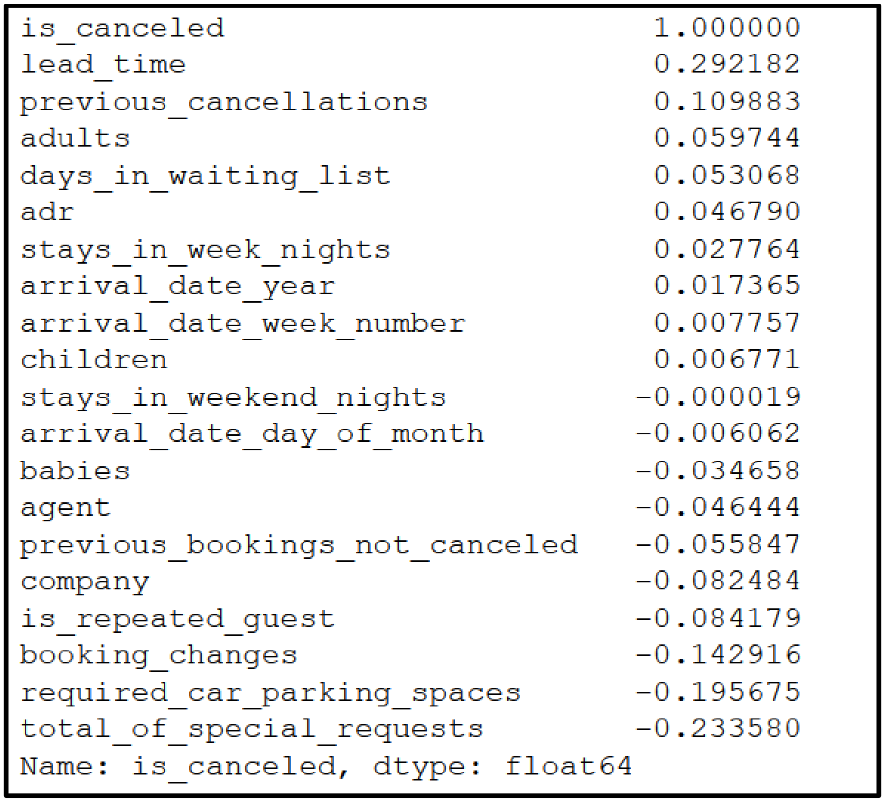
Fig 1: Columns with null values

Exploratory Data Analysis

Exploratory Data Analysis is utilized by information researchers to break down and research informational indexes, and sum up their fundamental attributes, frequently utilizing information representation strategies. It decides how best to control information sources to find the solutions you need, making it simpler for information researchers to find designs, spot abnormalities, test a speculation, or check presumptions.

For this project, we had quite a lot of data and needed to gain insights about various reasons as to why are the cancellations happening. We started our analysis with basics such as determining the correlation between the variables with respect to one variable.

1. Correlation



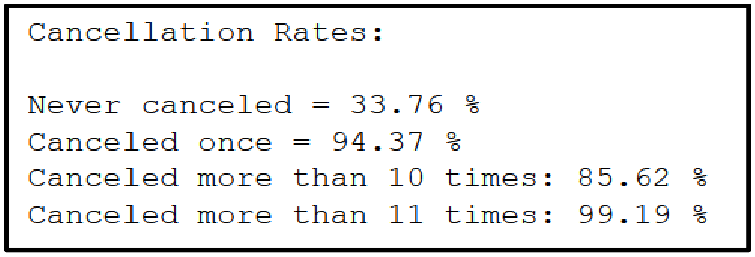
```
is_canceled                         1.000000
lead_time                           0.292182
previous_cancellations              0.109883
adults                              0.059744
days_in_waiting_list                0.053068
adr                                 0.046790
stays_in_week_nights                0.027764
arrival_date_year                   0.017365
arrival_date_week_number            0.007757
children                            0.006771
stays_in_weekend_nights            -0.000019
arrival_date_day_of_month          -0.006062
babies                             -0.034658
agent                              -0.046444
previous_bookings_not_canceled     -0.055847
company                            -0.082484
is_repeated_guest                  -0.084179
booking_changes                    -0.142916
required_car_parking_spaces        -0.195675
total_of_special_requests          -0.233580
Name: is_canceled, dtype: float64
```

Fig 2: Correlation in terms of “is_Cancelled” variable

From the above figure, we can see that there is no significant correlation between these variables. Although, Cancellation is somewhat related positively to the lead_time and negatively to the total_number_of_special_requests.

2. Cancellation Rates

Further, we determined the cancellation rates of customers.



```
Cancellation Rates:

Never canceled = 33.76 %
Canceled once = 94.37 %
Canceled more than 10 times: 85.62 %
Canceled more than 11 times: 99.19 %
```

Fig 3: Cancellation Rates

From the above figure, we can see that only 33.76% of the customers never cancel their bookings. The concerning number is that of the canceled more than 11 times, which is 99.19%. We can draw a conjecture from this and say that the customers, who might be canceling so many times, may just never end up making a booking. We can develop a cancellation policy keeping in mind to limit the number of times a customer can modify or cancel their bookings.

3. Cancellation trends

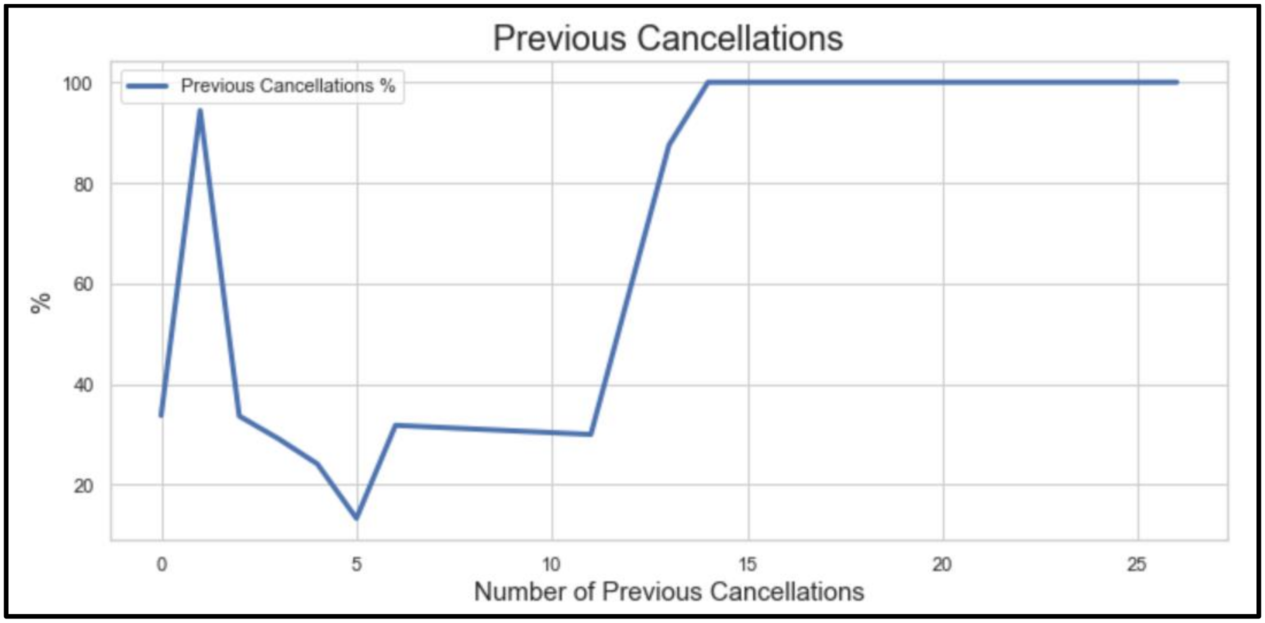


Fig 4: Cancellation trends

The graph above demonstrates the cancellation rate based on previous cancellations. It shows that more the customer has cancelled bookings in the past; chances of them cancelling the booking again increase drastically.

4. Cancellation by lead times

Further in our exploratory data analysis, we determined the cancellation rate based on the lead days as shown in the figure below. Earlier the customer cancels, more refund they get. Hence, the company needs to know how much amount is to be refunded in order to make a profit or engage a customer. Higher the amount charged for cancellation lesser chances will be there that the customer cancels.

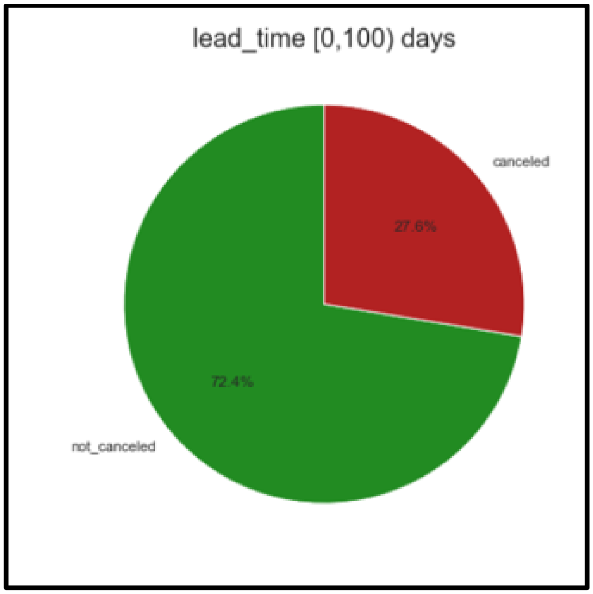


Fig 5: Cancellation with 0-100 lead days

We see that 27.6% of the cancellations are happening around 0-100 days before the check-in day.

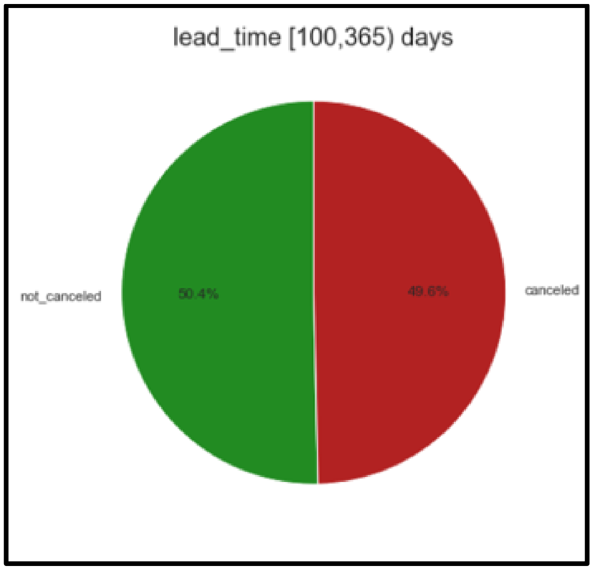


Fig 6: Cancellation with 100-365 days

We see that 49.6% of the cancellations are happening around 100-365 days before the check-in day, which is considerably more than the cancellation rate with 0-100 lead days.

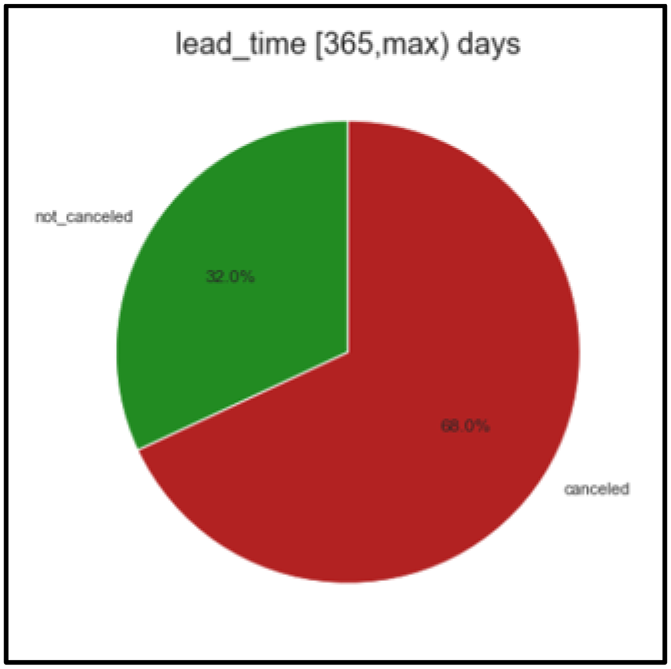


Fig 7: Cancellation with more than 365 days

When there is more than a year to booking, maximum cancellations happen with the cancellation rate being 68%.

After further analysis, we concluded that most bookings occur about 5 days prior to arrival. When the lead-time is larger, the chances for cancellation increase. The amount of bookings are steady overall between 20-100 days, then drops.

5. Cancellation based on special requests

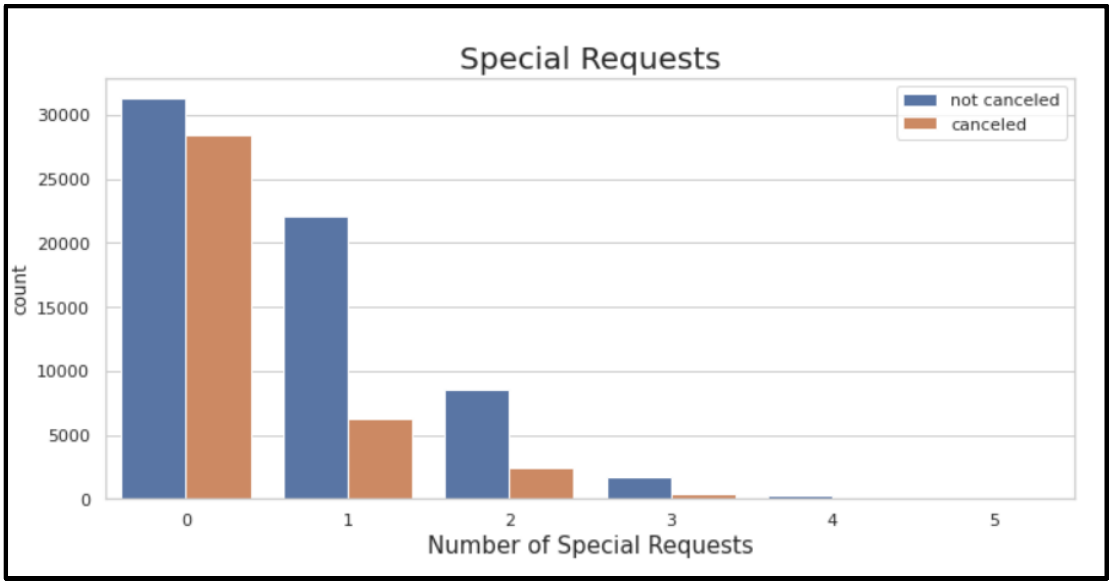


Fig 8: Cancellation based on special requests

If the customer is making more special requests, the chances of cancelling the booking are very less. Chances of cancellation are very high if there are no special requests made by the customer.

6. Cancellation based on Parking Space

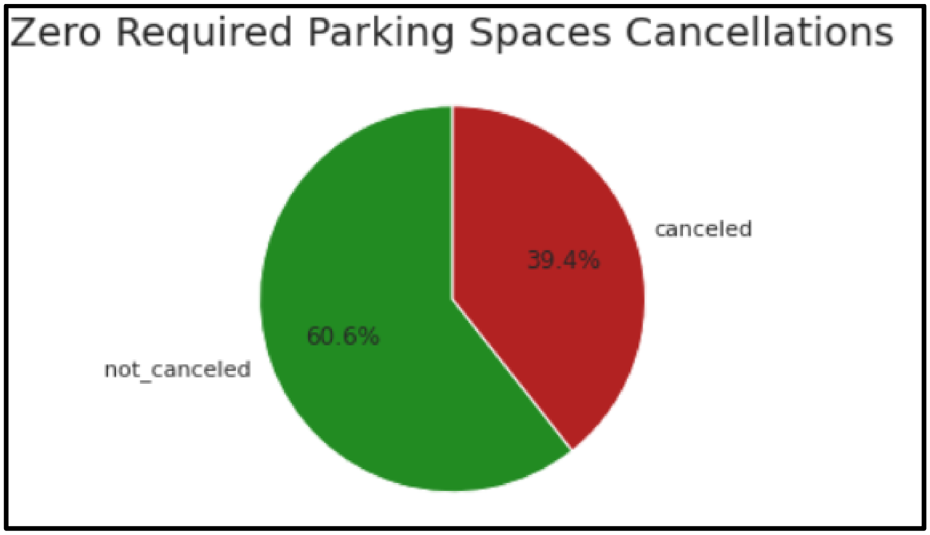


Fig 9: Cancellation based on Parking Spaces

If the customer was making any special request for the parking space reservations for their cars, then chances of them cancelling was only 39.4%. This combined with other factors can be accounted to determine the chances for cancellation of that customer and give them offers to not cancel booking.

Data Modeling

Data modeling is the way toward creating an unmistakable chart of connections between different kinds of data that are to be included in data set. One of the objectives of information demonstration is to make the most proficient technique for putting away data while accommodating total access and revealing.

We used multiple data modeling techniques based on the objectives we were trying to target. Firstly, we began with using K-NN algorithm to predict the reservation status and then used Random Forest classifier to predict the bookings. We compared the results from both of these algorithms to determine which one gives us better results. The aim was to choose the results from better performing algorithm for further analysis. The description of the models is given below.

K-NN

The KNN algorithm functions on the assumption that things, which are similar, are closer to each other in proximity. In other words, similar things are near to each other. The KNN algorithm hinges on the assumption of similar things being near and that in a way makes algorithm extremely useful. KNN calculates the distance between various data points using different formula. The most commonly used is Euclidean distance. The implementation of algorithm is as follows:

1. Load the data for the model
2. Initialize the value of “k” to chosen number of neighbors
3. For each point in the data
    a. Calculate the distance between the query example and the current example from the data.
    b. Add the distance and the index of the example to an ordered collection
4. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances
5. Pick the first K entries from the sorted collection

6. Get the labels of the selected K entries
7. Return the mode of the K labels

Random Forest

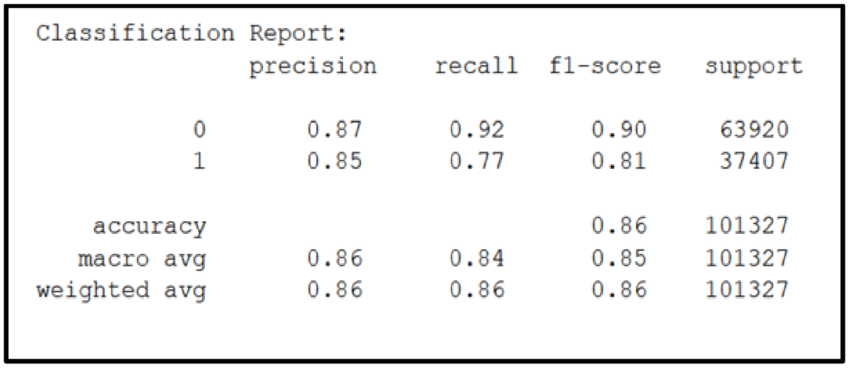
The random forest is an ensemble approach. It can also be thought of as a form of nearest neighbor predictor. Many a times to improve the performance of model, ensembles are used. They use the divide-and-conquer approach for that. Ensemble is built on the principle that many weak learners together can form a strong leaner. A classifier by itself is a weak learner, while all together are strong learner. Random forest classifier starts with Decision tree, which is weak learner. Decision trees take input at top and then traverse down the tree. As it traverses down, the data is bucketed into smaller sets. Each individual tree in the random forest provides a class prediction and the class with the most votes becomes model's prediction.

Discriminant Analysis

Discriminant Analysis is a technique, which aids the researcher to study difference between two or more group of objects with respect to several variations simultaneously. In this project, since the dataset is large, predictions are done using Python. A confusion matrix is extracted from there and then it is combined with average tariff of booking the hotel per night for further cost analysis. If the model correctly predicts the reservation status, we can determine if a cost is incurred or profit is made. Wrong predictions can lead to incurring cost as hotel staff may work towards customer retention whereas the customer is going to cancel booking despite that. However, if the prediction is correct, correct strategies can be implemented to retain the customer. Moreover, even if the customer is not going to cancel booking, hotel can think of more strategies to gain additional revenue from these customers based on their special requests or room types booked. Hence, by having a value to put on the predictions, we can determine the benefits of the model.

Model Results

After running the K-NN and Random Forest classifiers, we got the results as shown in images below.

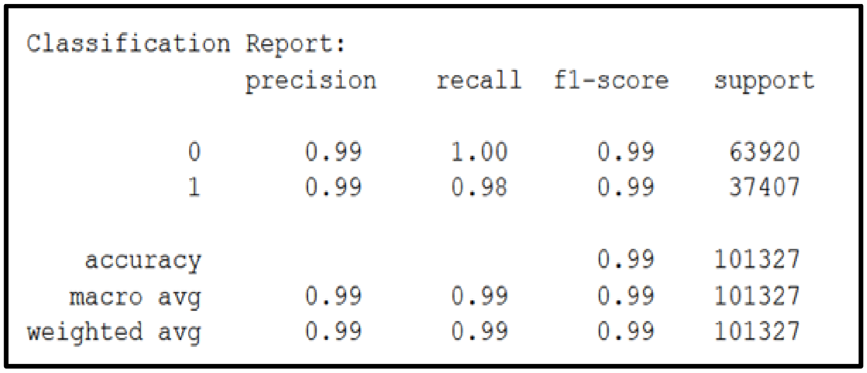


```
Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.92      0.90     63920
           1       0.85      0.77      0.81     37407

    accuracy                           0.86    101327
   macro avg       0.86      0.84      0.85    101327
weighted avg       0.86      0.86      0.86    101327
```

Fig 10: Classification report for K-NN model



```
Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99     63920
           1       0.99      0.98      0.99     37407

    accuracy                           0.99    101327
   macro avg       0.99      0.99      0.99    101327
weighted avg       0.99      0.99      0.99    101327
```

Fig 11: Classification report for Random Forest model

As we can see from the images above, Random forest classifier performs very well by giving us a precision of 0.99 as compared to K-NN, which only provides us with a precision of 0.86.

Evaluating Model Results

Confusion Matrix

Confusion matrix is used to describe the performance of a classification model on a set of data for which the true values are known. It consists of four concepts:

True Positive (TP) - These are data points which we correctly predicted to be yes

True Negative (TN) - These are data points which we correctly predicted to be no

False Positive (FP) – These are the data points we predicted to be a yes, but they are actually no.

False Negative (FN) - These are the data points we predicted to be a no, but they are actually yes.

From the models used in this project, I got the following results:

- Recall rate is 99%. This shows that the predictions about which customer will actually cancel the booking have been correctly predicted.
- Precision is also 99%, which means model correctly predicts actual cancelled bookings overall.

Discriminant Analysis

Based on the research of tariff rates in hotels during various seasons (Peak seasons and Off-seasons), we set the price for rooms as shown below:

Month	Tariff
January	$ 250.00
February	$ 118.00
March	$ 195.00
April	$ 215.00
May	$ 250.00
June	$ 215.00
July	$ 206.00
August	$ 188.00
September	$ 197.00
October	$ 197.00
November	$ 400.00
December	$ 500.00

Table 1: Tariff rates by month

An average of these rates were used for obtaining the Return Matrix:

	Predicted Positive	Predicted Negative
Actual Positive	$ 244.25	$ -
Actual Negative	$ (244.25)	$ -

Table 2: Return Matrix

Confusion matrix obtained from the Random Forest classifier:

	Not Cancelled	Cancelled	Total
Not Cancelled	63607	313	63920
Cancelled	682	36725	37407

Table 3: Confusion Matrix

Percent correct classification: 99.02%

Probability of classification of an Individual:

(Actual along side, predicted along top)

	Yes	No
Yes	0.628	0.003
No	0.007	0.362
Total	0.634	0.366

Table 4: Probability of classification of Individual

Expected Return can be calculated as shown below:

$$\left(\left(\frac{\text{Probability (predicted yes, actual yes)}}{\text{Sum Probability (predicted yes)}} \times \text{Average Return}\right) \times \text{True Positive}\right) + \left(\left(\frac{\text{Probability (predicted yes, actual no)}}{\text{Sum Probability (predicted yes)}} \times \text{Average Cost}\right) \times \text{False Negative}\right)$$

$$= \left(\left(\frac{0.628}{0.634} \times 244.25\right) \times 63607\right) + \left(\left(\frac{0.007}{0.634} \times (-244.25)\right) \times 682\right)$$

= $15,369,873.81

Conclusion

From the analysis, I was able to analyze and conclude the cost of unforeseen cancellations and not just predict it. The hotel was able to make a profit of a little above 15 million in time duration of two years. However, this amount can further be increased by using the model in this study and further take the following actions to reduce their cancellations:

- When the lead time is more, the cancellations of bookings increase. A possible reason for this might be that the customer may forget about the booking completely. If that is the reason, hotel can send monthly or weekly reminders leading up to the check-in day.
- We also noticed that when customer has more special requests, i.e. more customized stay planned at hotel, the chances of cancellation are very less. This is very interesting observation as the Hotels can use this insight, consider to provide a tailored experience to their customers, and reduce the chances of cancellation further.

Future Scope of this project can be extrapolating the results to a similar study or to other industries such as airline bookings or railway ticket bookings. Overall, for this project, hotels should consider investing more in increasing their customer engagement and retention strategies.

References

https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

https://towardsdatascience.com/understanding-random-forest-58381e0602d2