Predicting Cancellations of Hotel Reservations

Riya Mittal (106120098)

Yash Singhvi (106120148)

Introduction:

In this hotel reservations project, we will develop a machine learning model that can forecast whether or not a certain reservation made by a user will be canceled.

In this project, we will commence with collecting and segregating pertinent datasets. Followed by basic and advanced data analysis, furthermore, applying the concepts of machine learning to build a functional and optimized machine learning model.

Some of the domains and platforms where this project can be applied are online traveling and reservation platforms such as MakeMyTrip, GoIbibo, EaseMyTrip, Airbnb, Oyo, Booking.com, Yatra, Expedia, Agoda, etc. All such platforms shall benefit from this model to predict and project the cancellation of reservations that have been made by the customers.

All of these platforms may anticipate an increase in income by reducing the probability of hotel cancellation with the use of this data science and machine learning initiative.

Literature Survey:

1] Predicting Hotel Bookings Cancellation with machine learning classification model.

https://ieeexplore.ieee.org/document/8260781

AUTHORS: Nuno Antonio, Ann de Almeida, Luis Nunes

PROS: If we have required data accurate prediction is possible.

CONS: Incapable of producing highly accurate results.

2] Application of Machine Learning in the Hotel Industry.

https://jaauth.journals.ekb.eg/article_108732_19b7e16e7240aa896469ddc21fcc46bf.pdf

AUTHORS: Dr. Eid Alotaibi

PROS: Machine Learning Algorithms provides simplicity.

CONS: Social influence makes it difficult to control the predictions.

3] Aspect based Sentiment Oriented Summarization of Hotel Reviews

https://www.sciencedirect.com/science/article/pii/S1877050917319439

AUTHORS: Akhtae, Nashez Zubair, Abhishek Kumae, Tameem Ahmad

PROS: Analyze information that ratings would overlook.

CONS: Requires high amount of data analyse and process.

4] Machine learning algorithms are implemented in the embedded systems with the help of APIs.

https://www.researchgate.net/publication/340648863_Machine_learning_algorithms_implmentation_into_embedded_systems_with_web_application_user_interface

AUTHORS: Kamil Židek, Ján Piteľ, Alexander Hošovský

PROS: The Artificial Neural Network (ANN) method used gives better results than others.

CONS: The algorithm used does not indicate the sufficient accuracy.

Design and Implementation:

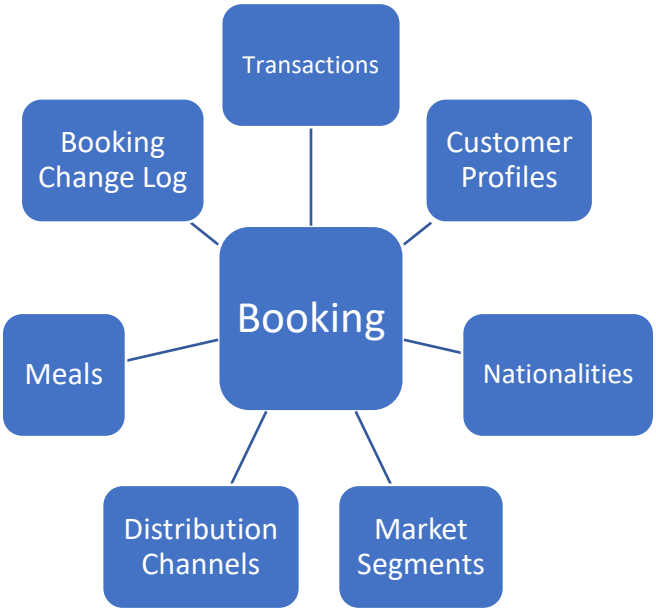


Fig 1. Diagram of Database Tables

Steps of Implementation:

1. Importing the data from the dataset (hotel_bookings.csv)
    - The following data set was imported from Kaggle.

2. Cleaning the data
    - We checked for null valued features.
    - Features that contained fewer null entries were imputed by replacing it with mode (i.e. COUNTRY).
    - Features that contained a lot of null entries were completely dropped from the data set (i.e. AGENT and COMPANY).
    - The records in the dataset that consisted of booking for 0 adults and 0 children, and 0 babies were dropped. (INVALID ENTRIES)

3. Classifying the nationalities of the guests
    - We calculated the mode of the guests that did not cancel their reservations with respect to their country and replaced the country with a numeric value.
    - This was further analysed with the total cumulative value of guests that turned up for their reservations.

4. Calculating how much guests pay for a room per night
    - Bookings which weren't canceled, were examined based on the hotel type. (i.e. RESORT and CITY HOTEL)
    - Followed by distribution based on room types and their Average Daily Rate.

5. Calculation of the busiest months in a year
    - We calculated the total rush in both the types of hotels by the number of guests that turned up each month.
    - This gave us the season for maximum bookings of the hotel and maximum footfall.

6. Calculation of the highest average daily rate classified by months
    - We compared the Average Daily Rate of bookings that were cancelled and not cancelled.

7. Analysis of whether bookings were made only for weekdays or for weekends or for both
    - The data was then further classified into bookings that were made only for weekdays, only weekends and both.
    - Some data that did not fit into these 3 categories were marked as UNDEFINED (entries that had more than 5 weekday bookings but 0 weekend bookings, etc.)
    - Data was sorted with respect to count and grouped by month.

8. Creating some more features
    - Four new features were added:
        1. *Family*: Booking that were made with at least one adult and at least one child or baby was defined as a family. This way three features were reduced to one, which helped refined the data set.
        2. *Total customers:* Sum of adults, children and babies were put into this feature.
        3. Total nights: The data from the reservations for weekdays and weekends were summed into this feature.
        4. *Deposit type: No Deposit, Non Refund, Refundable* were mapped to 0, 1 and 0 respectively to help analyse the data better.

9. Feature encoding of data
    - We converted all different categorical data types to numeric values.
    - This was achieved by using mean encoding. Mean encoding represents a probability of your target variable, conditional on each value of the feature.

10. Handling the outliers in the dataset
    - Generally, normal and gaussian distribution is best suited for an ML model.
    - To do so, we must handle the outliers which disrupt the distribution of the data.
    - To handle the negative entries in some records, we used log(x+1) instead of log(x).
    - We further dropped all records that had null values after applying logarithmic function.

11. Selecting important features using co-relation & univariate analysis
    - If there is a high correlation between two features, it is OVERFITTING and is not suitable for data analysis.
    - If there is a low correlation between two features, it is UNDERFITTING and is not suitable for data analysis.
    - So, we drop all features that had OVERFITTING or UNDERFITTING with *is_canceled* feature of the data set.

12. Finding important features for model building
    - The features were classified as DEPENDENT and INDEPENDENT using alpha values in lasso learning model.
    - The bigger the alpha value the fewer features that will be selected.

13. Building a machine learning model by splitting the dataset into training and testing data
    - We separated the entire data set into training and testing models.
    - We used 75% of the data to train and 25% to test the model accuracy.

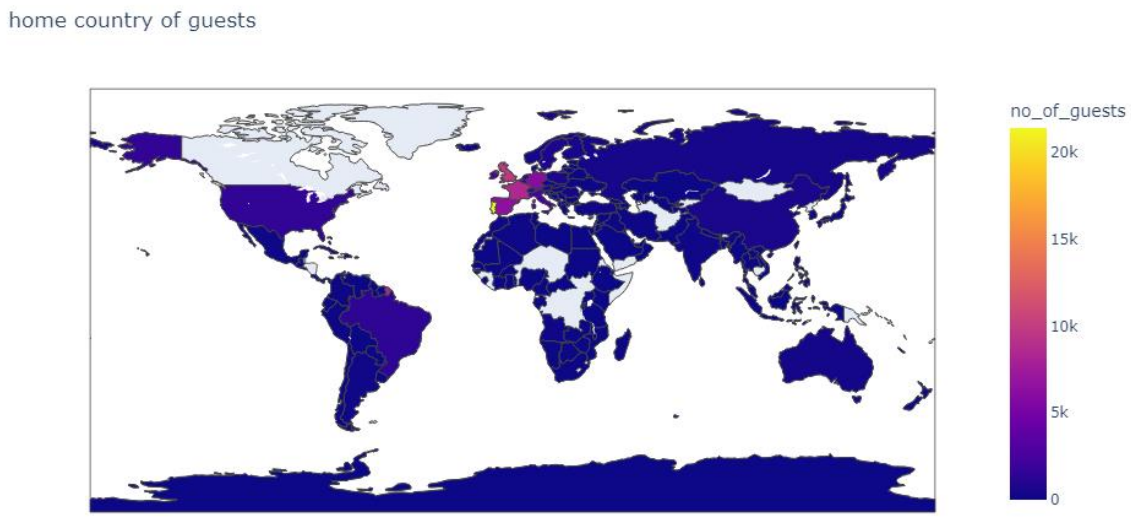
14. Cross-validating the model
    - Cross-validation is a statistical technique for assessing and contrasting learning algorithms that involves splitting the data into two sections: one for learning or training a model and the other for model validation.
    - We applied a 10-fold cross-validation model.

15. Applying multiple algorithms.
    - We applied several algorithms and calculated their accuracy percentage along with the confusion matrix.
    - Algorithms applied:
        1. *Logistic Regression*
        2. *Naïve Bayes*
        3. *K Nearest Neighbours (KNN)*
        4. *Decision Tree*
        5. *Random Forest*

16. Manually checking our model.
    - Random entries from the dataset were chosen to test the model and cross-checked with the *is_canceled* feature.

Result and Discussions:

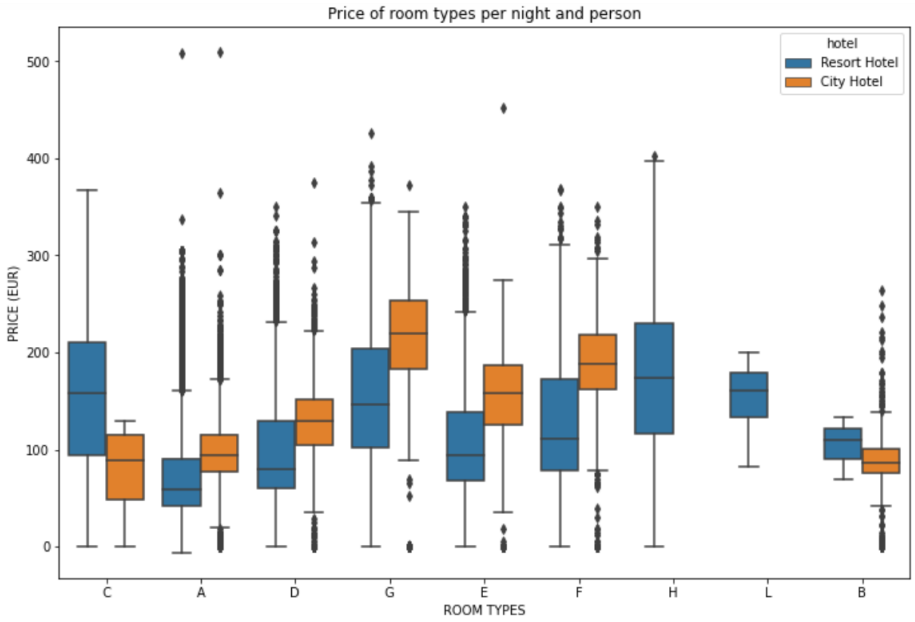
Nationalities of the guests:



Conclusion:

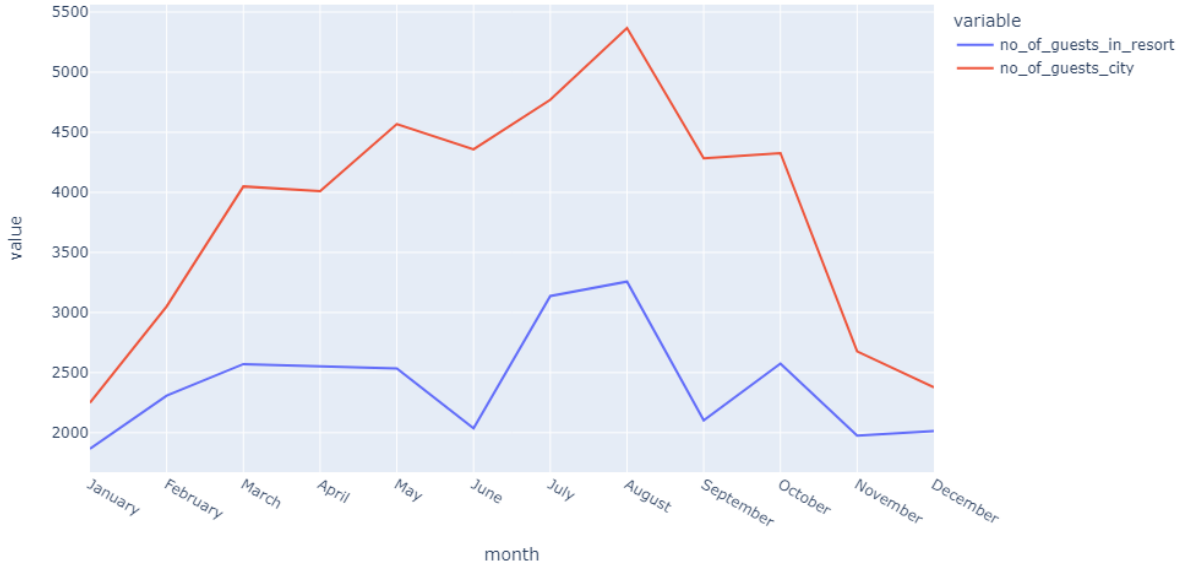
People from all over the world are staying in these two hotels. Most guests are from Portugal and other countries in Europe.

Prices of room types per night and person:



This figure shows the average price per room and the standard deviation, depending on its type. Note that due to data anonymization, rooms with the same type of letter may not necessarily be the same across hotels.

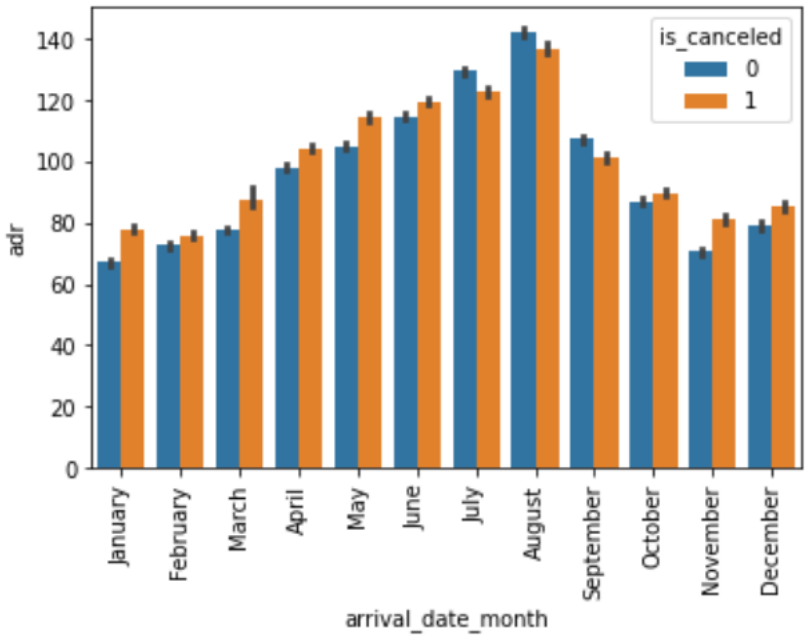
Rush of hotels in various months:



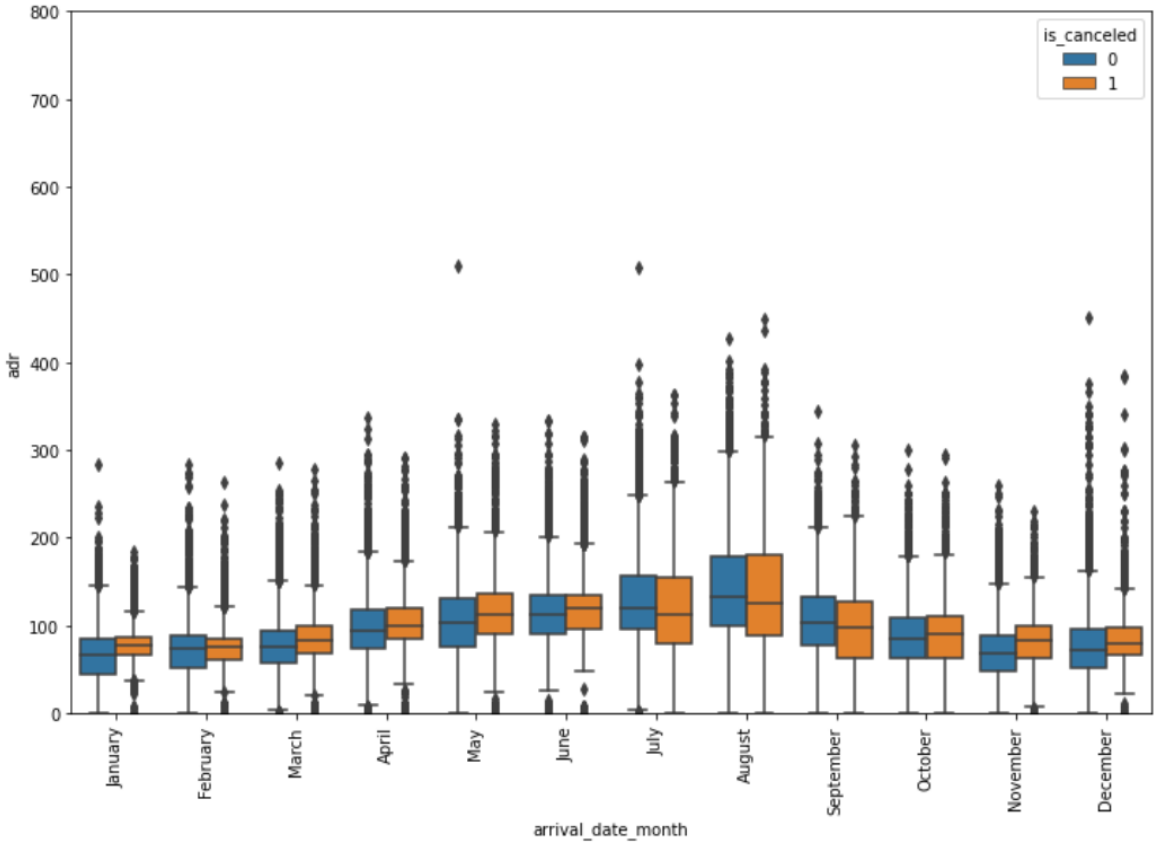
Conclusion:

This figure clearly shows that the rush in the Resort hotels are much higher during the summer. The rush in the city hotel varies less and is most expensive during spring and autumn.

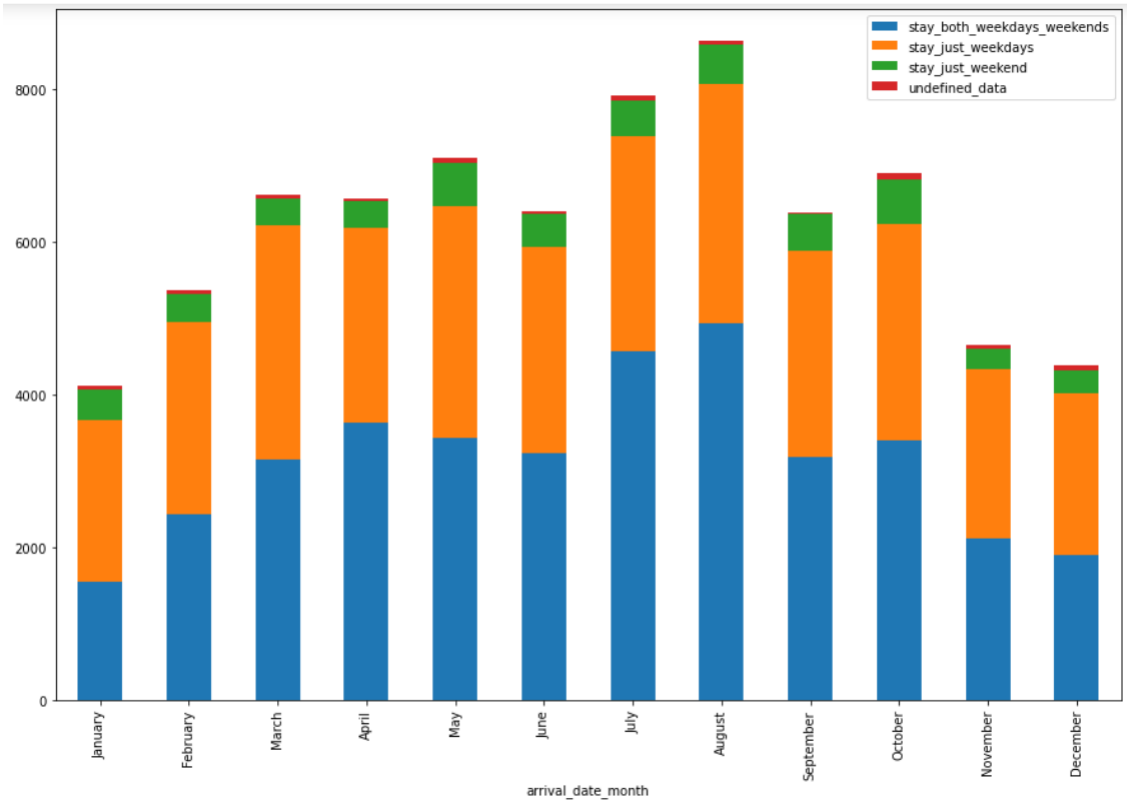
Average Daily Rate of bookings that were cancelled and not cancelled:



Checking for outliers and quadrants:



Stay for weekdays and weekends:



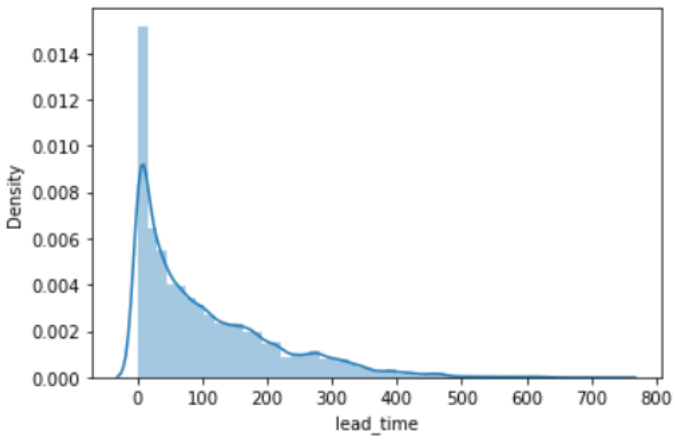
Conclusion:

Most bookings were made both for weekends and weekdays.

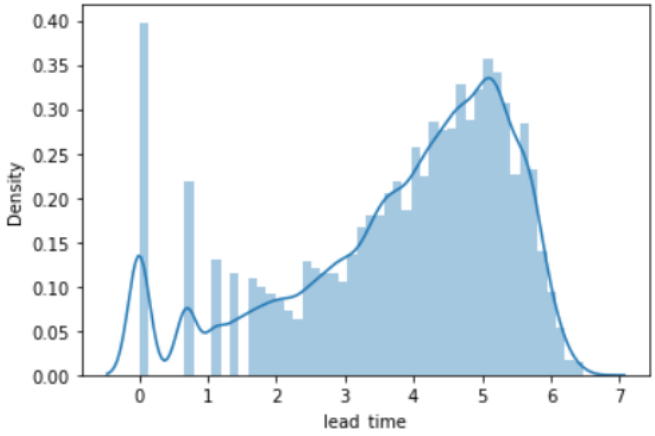
Handling Outliers:

a) *Lead time:*

Initial distribution:

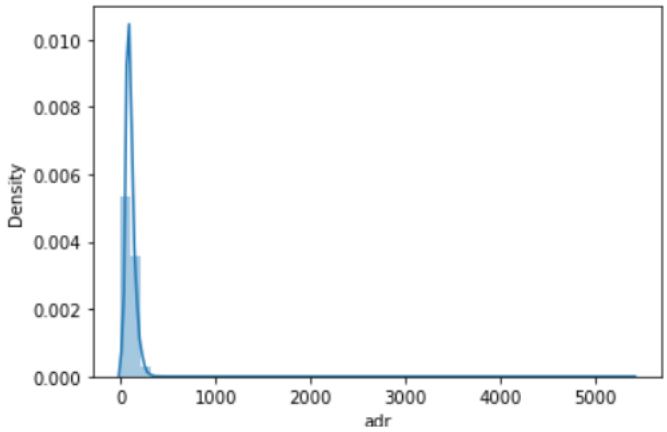


After applying logarithmic function:

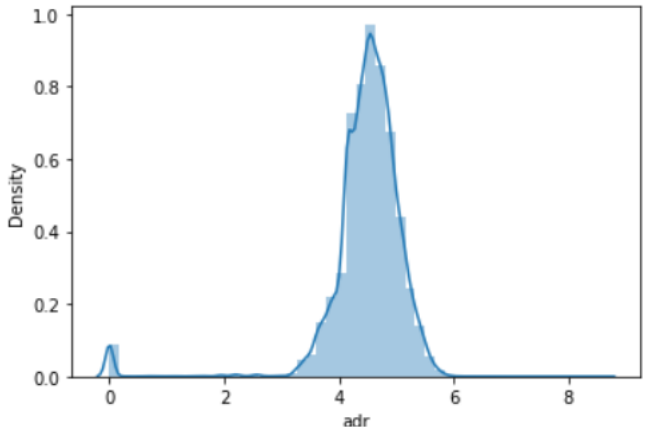


b) *Average Daily Rate:*

Initial distribution:

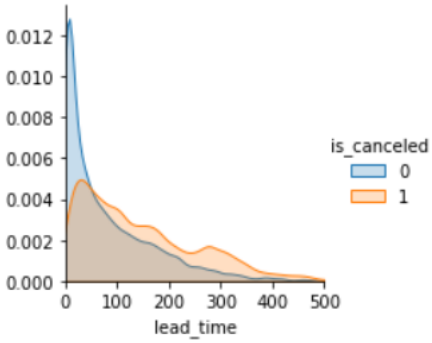


After applying ln(x+1) function



These null entries were then removed from the dataset.

Co-relation:



After applying various models:

```
LogisticRegression:
Confusion Matrix:
[[17464  1312]
 [ 4663  6364]]
Accuracy Score:
0.7995168271650505


Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.93      0.85     18776
           1       0.83      0.58      0.68     11027

    accuracy                           0.80     29803
   macro avg       0.81      0.75      0.77     29803
weighted avg       0.80      0.80      0.79     29803



Naive_bayes:
Confusion Matrix:
[[ 6674 12102]
 [  662 10365]]
Accuracy Score:
0.5717209676878167


Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.36      0.51     18776
           1       0.46      0.94      0.62     11027

    accuracy                           0.57     29803
   macro avg       0.69      0.65      0.57     29803
weighted avg       0.74      0.57      0.55     29803



Random Forest:
Confusion Matrix:
[[17230  1546]
 [ 2780  8247]]
Accuracy Score:
```

```
0.8548468275005872


Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.92      0.89     18776
           1       0.84      0.75      0.79     11027

    accuracy                           0.85     29803
   macro avg       0.85      0.83      0.84     29803
weighted avg       0.85      0.85      0.85     29803



Decision_tree:
Confusion Matrix:
[[15976  2800]
 [ 2675  8352]]
Accuracy Score:
0.8162936617119082


Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.85      0.85     18776
           1       0.75      0.76      0.75     11027

    accuracy                           0.82     29803
   macro avg       0.80      0.80      0.80     29803
weighted avg       0.82      0.82      0.82     29803



KNN:
Confusion Matrix:
[[16847  1929]
 [ 3409  7618]]
Accuracy Score:
0.8208905143777472


Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.90      0.86     18776
           1       0.80      0.69      0.74     11027

    accuracy                           0.82     29803
   macro avg       0.81      0.79      0.80     29803
weighted avg       0.82      0.82      0.82     29803
```

Conclusion and Future works:

We can conclude with our study, that given 12 features such as *'country', 'lead_time', 'previous_cancellations', 'previous_bookings_not_canceled', 'booking_changes', 'days_in_waiting_list', 'adr', 'required_car_parking_spaces', 'total_of_special_requests', 'total_customer', 'total_nights', 'deposit_given',* we can predict the cancellation of hotel reservations accurately 85% of the times using *Random Forest Algorithm.*

The advantages of having such a system are clearly demonstrated, even if this study is still in its early stages and these are only preliminary findings. The system demonstrates how a machine learning system to anticipate hotel booking cancellations may be created, implemented, and how it influences business, both theoretically and practically when seen as a machine learning prototype that addresses an unsolved problem in a novel and inventive way.

From a commercial perspective, it is generally acknowledged that a hotel lacks the means to get in touch with each visitor in advance of their visit. However, as this study showed, visitors who are contacted by hotels cancel reservations significantly less frequently than those who are not. Another well-known reality is that hotels do not have the contact information for every potential guest, and certain visitors are less receptive to deals and discounts (e.g. corporate guests). The incorporation of such a machine learning predictive model in a reservation management system could still assist hoteliers in decreasing the number of reservations to be contacted and, as a result, lead to reductions cancellation rates, at restricted prices. As a result, hoteliers cannot expect to contact all guests recognised as “likely to cancel”.

Future study on these models might benefit from adding information from more data sources on elements like rival prices, social media reputation, and weather, among other things, that influence customers' decisions to book or cancel. A feature that determines if an action to prevent cancellation was previously done on the booking has the potential to increase model accuracy by taking into account the influence actions can have on consumers' decisions not to cancel.

References:

[1] N. Antonio, A. Almeida, L. Nunes, Predicting hotel bookings cancellation with a machine learning classification model, in:

Proceedings of the 16th IEEE International Conference Machine Learning Application, IEEE, Cancun, Mexicopp. 1049–1054.

doi: 10.1109/ICMLA.2017.00-11, 2017.

[2] International Civil Aviation Organization, Guidelines on Passenger Name Record (PNR) data, (2010). ⟨ https://www.iata.org/iata/passenger-data-toolkit/assets/doc_library/04-pnr/New%20Doc%209944%201st%20Edition%20PNR.pdf ⟩ (accessed 17 February 2016).

[3] D. Abbott, Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst, Wiley, Indianapolis, IN, USA, 2014.

[4] Microsoft, SQL Server Management Studio (SSMS), (2017).

⟨ https://docs.microsoft.com/en-us/sql/ssms/sql-server-management-studio-ssms ⟩ (accessed 24 March 2018).

[5] American Hotel & Lodging Association, Uniform System of Accounts for the Lodging Industry, 11th Revised edition.

Educational Institute, New York, 2014.

[6] International Standards Organization, ISO country codes 3166-3:2013,

⟨ https://www.iso.org/obp/ui/#iso:std:iso:3166:-3:ed-2:v1:en,fr ⟩ (accessed 24 March 2018), 2013.

[7] A. McNamara, E.A. de la Rubia, H. Zhu, S. Ellis, M. Quinn, skimr: Compact and flexible summaries of data. R package version

1.0.1⟨ https://CRAN.R-project.org/package=skimr ⟩, 2018.

[8] M. Tennekes, E. de Jonge, tabplot: Tableplot, a visualization of large datasets

⟨ https://CRAN.R-project.org/package=tabplot ⟩,2017.

[9] Application of Machine Learning in the Hotel Industry: A Critical Review, Dr. Eid Tourism, Archaeology Department, College of Arts, University of Hail, P.O.Box 2440 Hail, Saudi Arabia, 2020.

[10] A Grouping Hotel Recommender System Based on Deep Learning and Sentiment Analysis, Fatemeh Abbasi, Ameneh Khadivar, Mohsen Yazdinejad, 2019.

[11] Modelling the cancellation behaviour of hotel guests, Martin Falk, Markku Vieru, 2018.

[12] A first attempt to address the problem of overbooking study programs, Karin hart, 2018.

[13] Aspect based Sentiment Oriented Summarization of Hotel Reviews, Akhtae, Nashez Zubair, Abhishek Kumae, Tameem Ahmad, 2017.

[14] Machine learning algorithms implementation into embedded systems with web application user interface , Kamil Židek, Ján Piteľ, Alexander Hošovský, 2017 IEEE 21st International Conference on Intelligent Engineering Systems (INES)2017.

[15] Machine Learning for Web Page Adaptation, Neetu Narwal, Dr. Sanjay Kumar Sharma, 2016.

[16] Predicting hotel booking cancellations to decrease uncertainty and increase revenue, Ana de Almeida ISCTE Instituto Universitário de Lisboa, 2016.

[17] Studying the cancellation behaviour of the guests, Markku Vieru, 2016.