**Predicting Hotel Booking Cancellation**

Submitted by:

AKANKSHA SINGH (230920525002)

SHREE TIWARI (230920525013)

Course-PGDABDA

Under the Guidance of:

Priti Bhardwaj



CDAC NOIDA

**Candidate Declaration**

I solely declare that the report titled “Predicting Hotel Booking Cancellation” is a bonafied record of work carried by us. Submitted partial fulfilment of requirement for the award of Post graduation diploma in Big Data Analytics under the guidance of Ms. Priti Bhardwaj The matter embodied in this report has not been submitted for the award of any other degree.

**Acknowledgement**

I would like to thank Ms. Priti Bhardwaj Ma’am for her help through this project. Her suggestions are the keys to the successful completion of this project and to understand the basic of analysis and design of algorithms which is the most important factor behind implementing efficient code. I would also like to thank Mr. Shivam Pandey Sir and Ms. Sidhidhatri Nayak for giving us suggestions to improve our project.

**Abstract**

In the 'Hotel Booking Prediction' project, the focus is on providing users with an accurate forecast of booking cancellations, allowing them to gain early insights into hotel booking-related matters. The impact of booking cancellations on revenue is substantial, influencing crucial decision-making in the hospitality industry. To mitigate this effect, the project employs a machine learning-based cancellation model as a key solution. By integrating data science tools and capabilities with human judgment, the project aims to showcase how predictive analysis can effectively address the challenge of booking cancellation forecasting.

Additionally, the project endeavours to offer users seeking accommodation in a specific hotel a comprehensive prediction and analysis. Through the implementation of various algorithms such as Logistic Regression, Random Forest, Decision Tree, and others, along with the use of Evaluation Matrices to categorize data, users can obtain predictions tailored to their desired level of accuracy. This approach not only benefits hotels by minimizing the impact of cancellations but also ensures a smoother experience for tourists, preventing issues related to room availability. Users or customers input specific information, enabling the model to make accurate predictions regarding cancellations.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sr. no.** | **Content** | **Page no.** |
| 1 | Introduction | 6-7 |
| 2 | Data collection | 8-9 |
| 3 | Data Visualization | 9 |
| 4 | Exploratory Data Analysis | 10-14 |
| 5 | Data Preparation | 14-16 |
| 6 | Modelling | 16-20 |
| 7 | Model Evolution and comparison | 21 |
| 8 | Conclusion | 22 |
| 9 | Recommendation | 23 |
| 10 | References | 24 |
| 11 | Bibliography | 24 |

1. **INTRODUCTION**

Hotels play a crucial role for travellers, whether local or international, as they journey from one destination to another. These establishments offer essential services such as parking, food, room service, and personalized offerings based on customer preferences. Hotels actively seek valuable feedback from guests to uphold their reputation in the city or area. The quality of services directly influences hotel bookings; poor services result in lower bookings, while excellent services lead to increased reservations.

In this predictive model, we assess the likelihood of hotel bookings based on various factors and also attempt to anticipate any special requests users might have, considering different features. The dataset employed in this project encompasses both international and local hotel data. Utilizing popular Machine Learning Algorithms including Decision Tree, Random Forest, Logistic Regression, among others, we aim to predict the probability of hotel booking cancellations. The model achieves a prediction accuracy of approximately 95%.

PROBLEM STATEMENT

The challenge is to utilize guest data and discern booking behaviour patterns through a Machine Learning approach to formulate an effective strategy for Hotel Booking Management.

The escalating trend of last-minute hotel cancellations poses a growing concern for both local and international tourists. Currently, customers lack information regarding the likelihood of cancellations when booking a hotel.

Solution: The proposed 'HOTEL BOOKING PREDICTION' seeks to empower users with insights into the cancellation percentages of hotels well in advance. Armed with this predictive information, users can make informed decisions about whether to proceed with their hotel/resort booking or not. Additionally, this tool enables hotel managers and revenue managers to proactively address bookings identified as "potentially going to be cancelled." The successful implementation of this predictive model is expected to significantly contribute to enhancing hotel revenue management strategies.

LITERATURE SURVEY

A. Machine Learning Classification for Predicting Hotel Booking Cancellations

In the contemporary world, hotel booking cancellations pose a common challenge that can result in substantial business losses. This report outlines the utilization of AI to discern cancellations, mitigating potential losses.

B. Escalating Cancellation Rates in the Hotel Industry

While the upward trajectory of the hotel industry is advantageous, it brings challenges like a rising rate of cancellations. Users often cancel bookings based on reviews, and instances of poor treatment by hotel owners adversely impact reputation and lead to cancellations. The contemporary trend shows a growing expectation for superior accommodation, and any shortcomings result in negative ratings. The cancellation percentage has been steadily increasing, with a survey indicating a rise from under 33% in 2014 to 40% in 2018. The COVID-19 pandemic further exacerbated this trend as changing circumstances and lockdown implementations prompted sudden changes in travel plans, leading to increased cancellations.

C. Aspect-Based Sentiment-Oriented Summarization of Hotel Reviews

Customer reviews significantly influence a hotel's image and revenue system, yet many travellers do not read them all. This system analyses reviews and feedback, collected from the hotel's website and categorized as classes. The model identifies overlooked information, takes necessary steps based on the analysis, and conducts emotional assessments. Hotels can then make informed decisions to enhance their services.

D. Critical Review of Machine Learning Applications in the Hotel Industry

The gradual integration of the IT industry into the hotel sector is a noteworthy development. Researchers are actively exploring the application of new artificial intelligence technologies and learning tools in the hotel industry. This study provides insights into the utilization of machine learning and associated technologies in the hotel and tourism sector, highlighting the current trends in this rapidly evolving field.

1. **Data collection**

The dataset utilized in this project has been sourced from Kaggle.com and originates from the article "Hotel Booking Demand Datasets" by Nuno Antonio, Ana Almeida, and Luis Nunes, published in Data in Brief, Volume 22, February 2019.

This comprehensive dataset encompasses information on hotels located in both local and international destinations, with a predominant focus on hotel data from Portugal. The dataset is structured around two distinct hotel types: Resort and City Hotel. In this context, City Hotel refers to data associated with urban hotels, while Resort Hotel pertains to establishments functioning as a combination of hotels and resorts.

The dataset incorporates four main categories of data points, as outlined below:

Market Segments: This includes information on the sources of booking generation, distinguishing between offline and online channels, along with relevant variables.

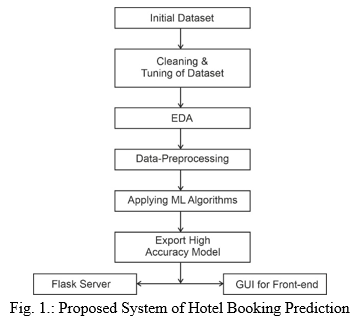
Hotel & Revenue: This category encompasses details such as the type of hotel (City or Resort Hotel), Average Daily Rate (ADR), and other revenue-related parameters.

Customer Related: Variables in this category describe the characteristics of the customer, offering insights into the type of stay.

Cancellation History: Information related to the cancellation history of bookings, indicating whether cancellations have occurred previously, among other relevant details.

This dataset provides a rich foundation for the analysis and prediction of hotel booking behaviour, allowing for a comprehensive exploration of various factors influencing the industry.

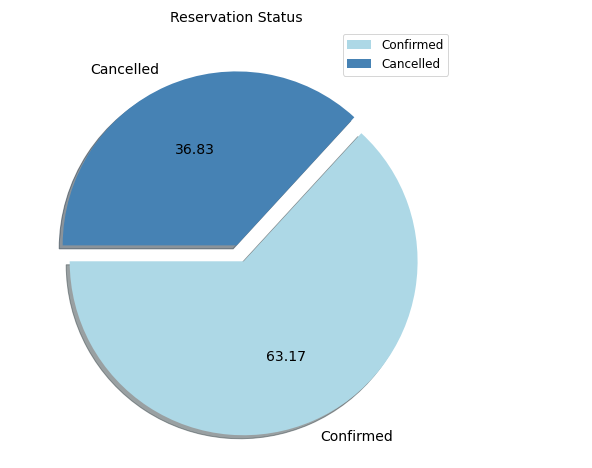
**PROPOSED SYSTEM**



1. **Data Visualization**

3.1 Cancellation percentage

According to the graph below, 63.17% of reservations were confirmed and 36.83% of the bookings were cancelled or guests didn't show up at the hotel.

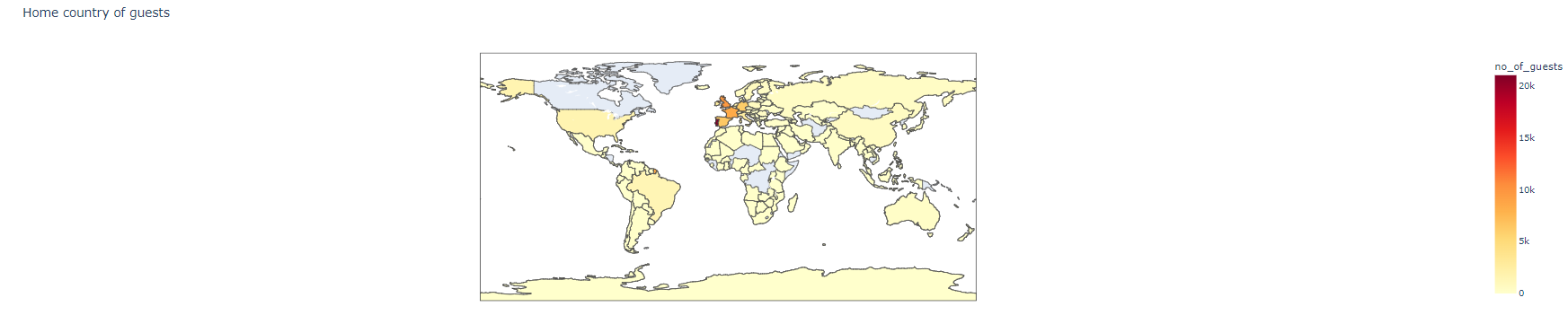


3.2 Table for MISSING OR UNDEFINED DATA

|  |  |
| --- | --- |
| **Feature** | **Number of Missing Features** |
| Company | 112593 |
| Agent | 16340 |
| Country | 488 |

1. **Exploratory Data Analysis**

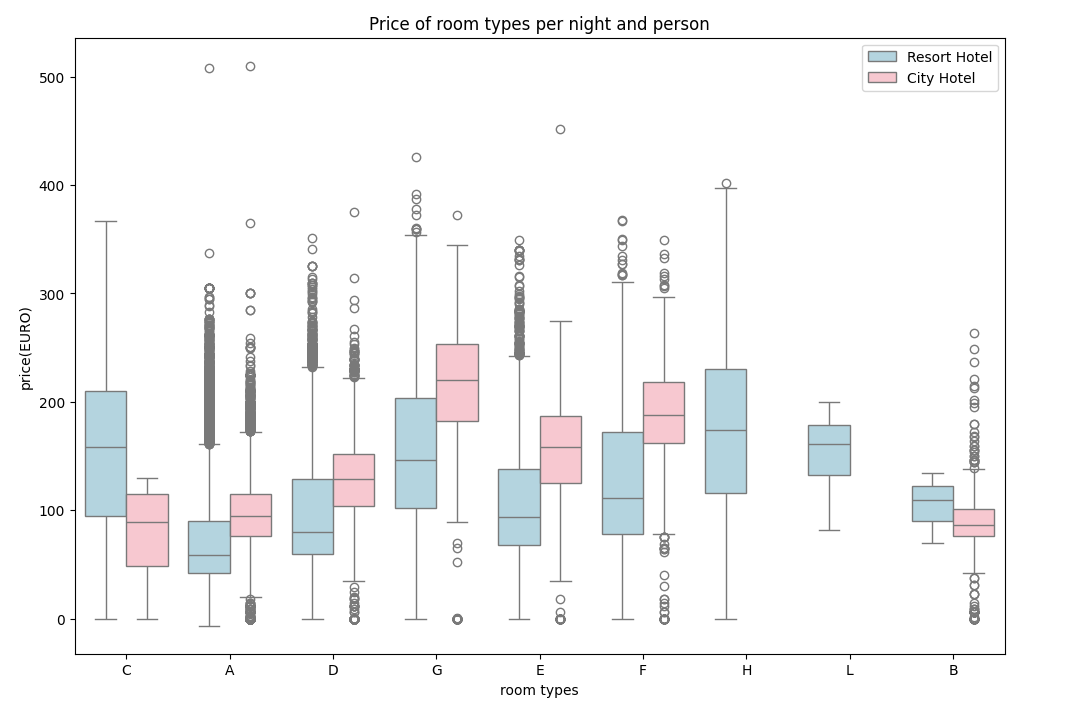
* **Where do the guests come from?**



Conclusion:

there are 80 to 85% of countries that have almost bewtween 0 to 5000 guests. And there are few countries ,which are exactly portugal and some European countries where we have maximum number of guest.

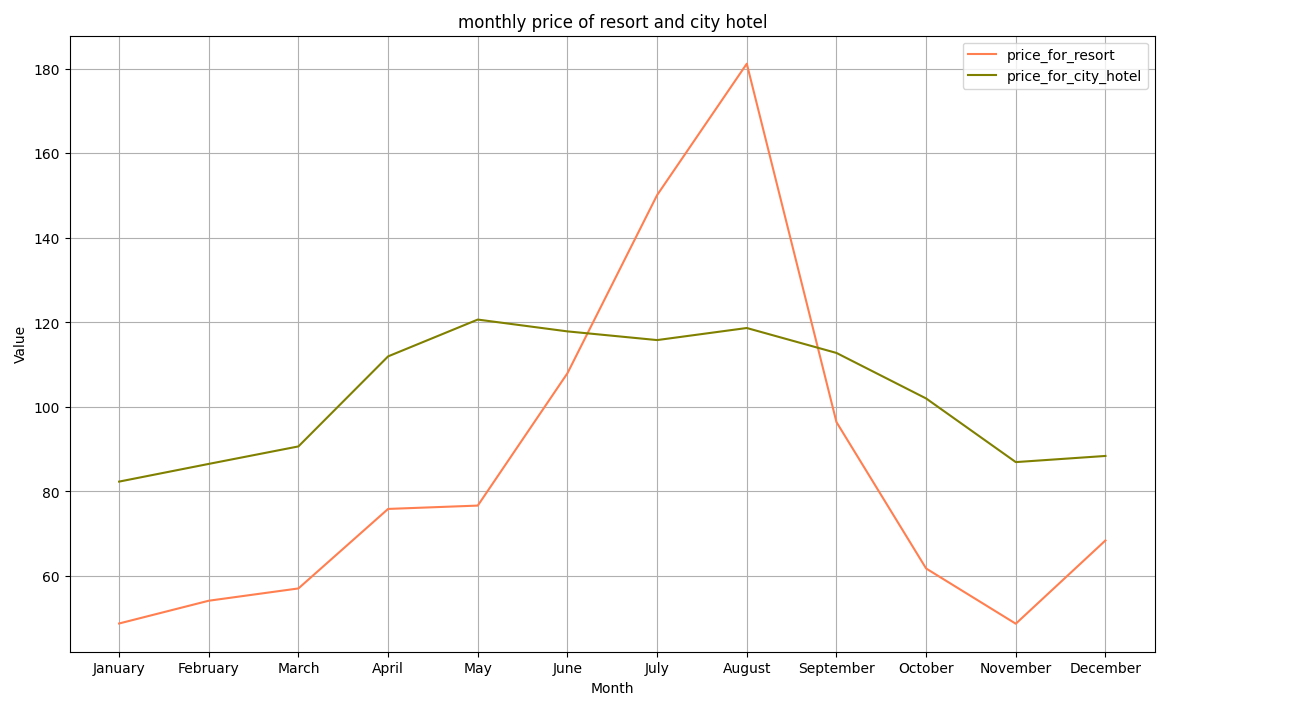
* **How much do guests pay for a room per night?**



CONCLUSION:

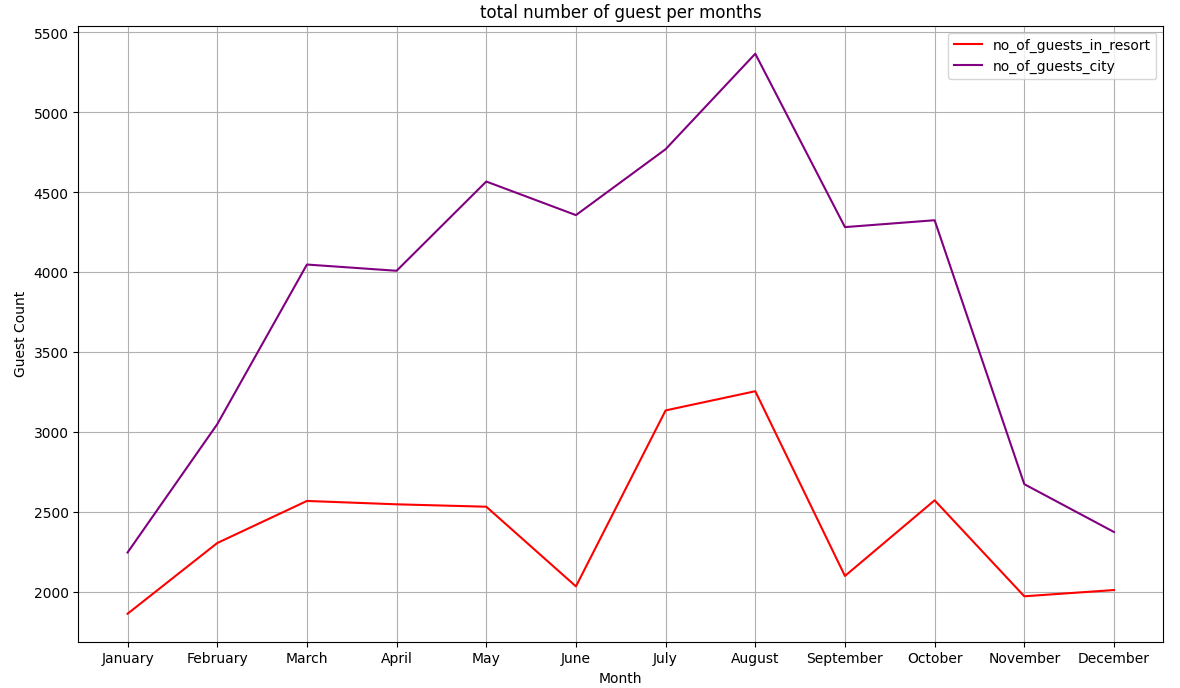
the best distribution with respect to City Hotel is almost with G type, whereas resort hotel, the best distribution of price is almost with H room type.

* **how does price per night vary over the year?**



Conclusion: it clearly shows that prices in the resort hotel are much higher during the summer and the prices for the city hotel will rise very less. it is most expensive during basically autumn.

* **which are the most busy month or in which months guests are high??**

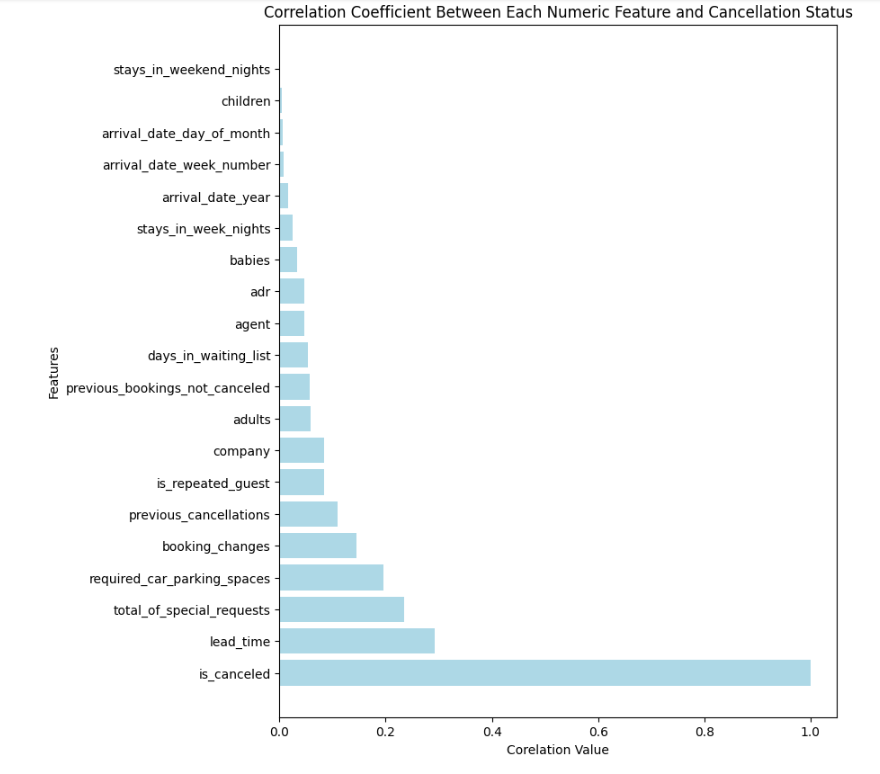


Conclusion:

The City Hotel has more guest during spring and autumn in July and August. There are less visitors you will observe, although prices are lower.

Whereas with respect to this resort hotel, Guest member for this resort hotel go down slightly from June to September. which is also when the prices is higher.

* **Visualizing correlation coefficients between features and cancellation:**

****

Interpretation:

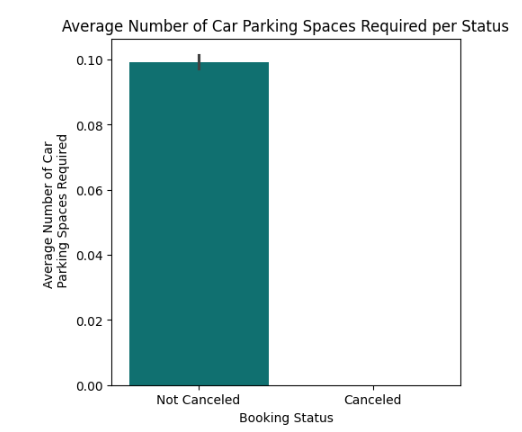
Lead time is the most highly correlated feature with whether or not a booking is cancelled. It makes sense that as the number of days between when the booking is made and the supposed arrival date increases, customers have more time to cancel the reservation.

the total number of special requests is the second feature with the strongest correlation to our cancellation target. As the number of special requests made increases, the likelihood that a booking is cancelled decreases. This suggests that engagement with the hotel prior to arrival and feeling like their needs are heard may make a customer less likely to cancel their reservation.

Related to special requests, the number of required car parking spaces is the third feature with the strongest correlation to our cancellation target. As the number of parking spaces requests increases, the likelihood that a booking is cancelled decreases.

a customer's prior history with the hotel (measured by the number of previous bookings not cancelled or whether or not a customer is a repeated guest) does not seem to be highly correlated with whether or not the current booking will be cancelled. On the other hand, a customer's prior history of cancellation (measured by the number of previous cancellations is more highly correlated with whether or not the current booking will be cancelled.

* **Visualizing the total number of requested parking spaces for cancelled and not cancelled bookings.**

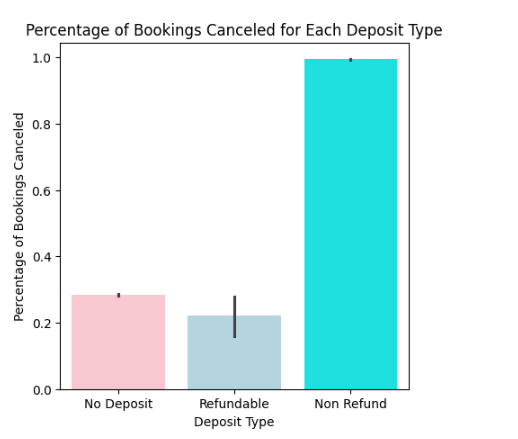


\*Interpretation: \*

On average, customers who do not cancel their bookings tend to require more parking spaces. Similarly to the number of special requests, it would make sense that the more a customer engages with the hotel (by putting in a request for a parking spot), the less likely they are to cancel.

From the hotel perspective, it is possible that not many hotels around have a parking. As a result, the need for a parking space would limit the customer in their hotel options and make them less likely to cancel.

* **Visualizing percentage of bookings canceled for each deposit type:**



\*Interpretation: \*

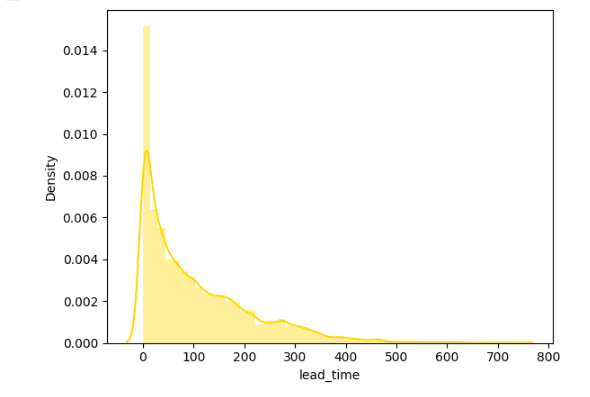
Surprisingly, customers who pay a non-refundable deposit have a much higher percentage of cancelled reservations.

1. **Data Preparation**

In this phase, the focus is on creating the final dataset that will be fed into the established model. Data preparation encompasses various activities aimed at constructing a dataset suitable for model utilization. Specifically, the emphasis lies on selecting features for the model and incorporating several feature engineering techniques. The ultimate feature set employed in the model comprises 24 features. However, six features are excluded to prevent data leakage.

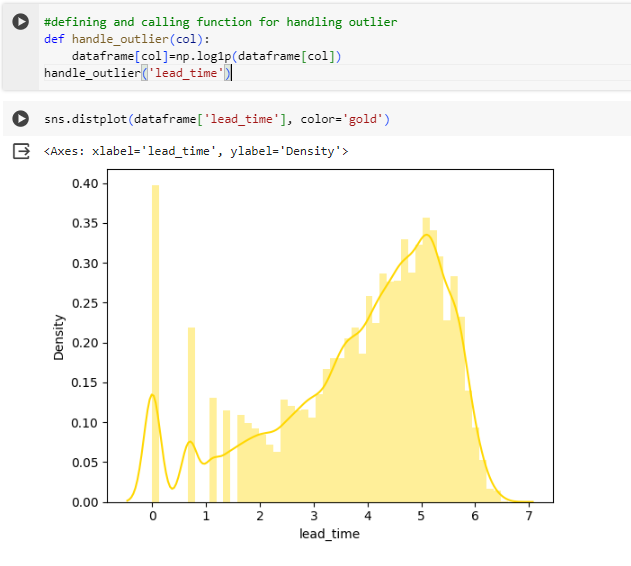
Additionally, three time-series features related to arrival time are transformed into a single feature named "day." This new feature represents the day's distance from January 1 to the booking time within the same year.

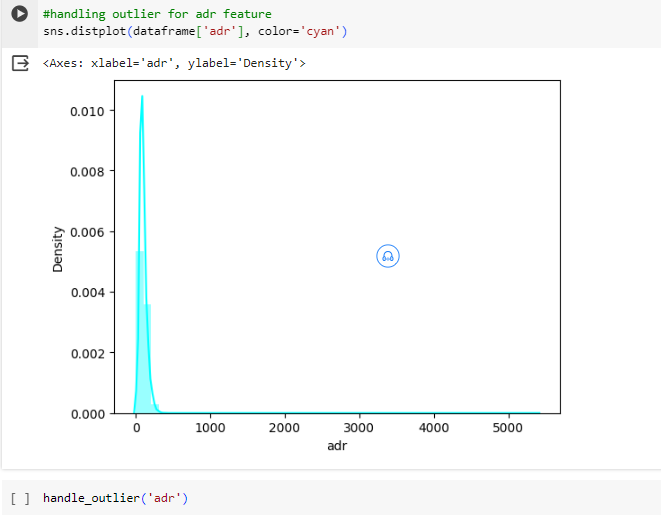
To handle categorical features like meal, market\_segment, distribution\_channel, deposit, agent, customer\_type, agent, and company, the mean encoding technique is applied. The only exception is the reserved\_room\_type feature, which is converted into an integer by assigning numerical values following alphabetical order. This meticulous data preparation ensures that the dataset is optimized for the model's effective utilization while addressing potential data leakage concerns. To handle outliers we used log method.

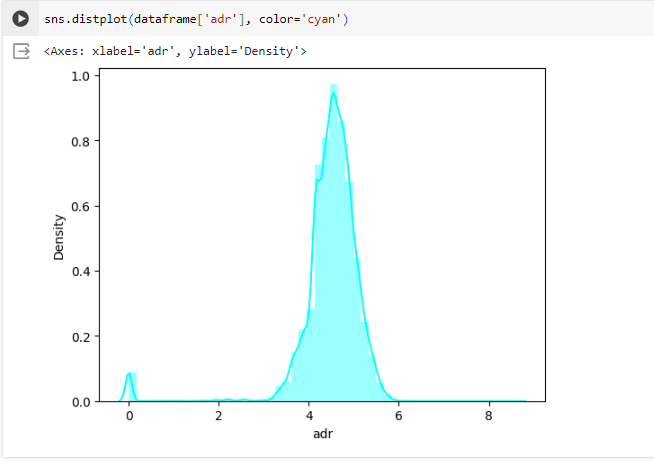


Conclusion:

the distribution of this features is little bit right skewed. we are going to replace the value of its original feature from it log value, so that the skeweness will get solved.







1. **Modelling**

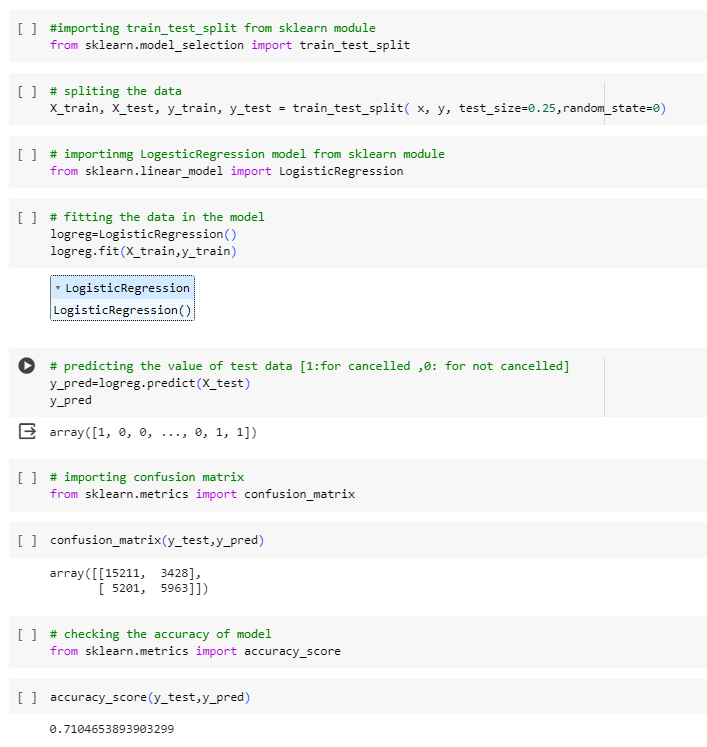
In this phase, the process involves the careful selection and implementation of machine learning models, aiming to identify the most effective one. The chosen models include tree-based algorithms such as Decision Tree (DT), Logistic Regression (LG), Naïve Bayes (NB) and Random Forest (RF). The decision to utilize tree-based algorithms in predicting hotel booking cancellations is grounded in their demonstrated effectiveness in prior studies.

These tree-based models have shown notable success in addressing prediction challenges, and we have selected them to leverage their strengths in handling the complexities of hotel booking cancellation predictions. The models selected for this study have proven track records in achieving accurate predictions and are anticipated to provide valuable insights into the hotel booking cancellation scenario.

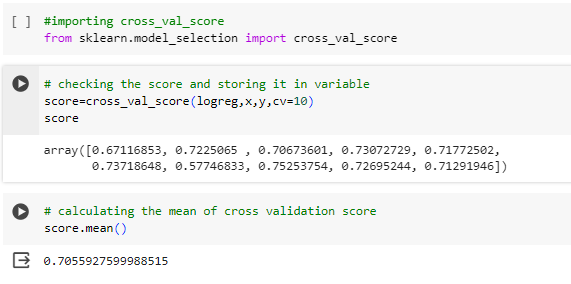
**Logistic Regression**

Logistic regression stands among the most widely used Supervised Machine Learning algorithms. It is designed to predict a categorical dependent variable by employing a given set of independent variables. The categorical dependent variable typically takes binary values such as Yes or No, 0 or 1, true or false, etc. Although logistic regression shares similarities with linear regression, its application is distinct. Linear regression addresses regression problems, while logistic regression is tailored for solving classification problems.

In contrast to fitting a regression line, logistic regression involves fitting a sigmoid S function. This sigmoid function outputs values between 0 and 1, representing the likelihood of a specific outcome. For instance, in weather prediction, logistic regression can assess the probability of rain based on given weather conditions. The sigmoid curve in logistic regression effectively models the probability distribution for binary classification scenarios.



this is not the correct accuracy of this model we need to do cross validation to get the correct the accuracy of model.



conclusion :

Since the logistic regression model does not provide predictive power as high as we would like, we will attempt other machine learning algorithm.

**Decision Tree Classification Algorithm**

The Decision Tree algorithm is a member of the non-parametric, supervised learning algorithms. It is versatile, capable of addressing both regression and classification problems. In the context of supervised learning, decision trees serve as powerful tools for making decisions based on input features. The algorithm works by recursively splitting the dataset into subsets based on the most significant features, creating a tree-like structure of decisions. This structure makes it easy to interpret and visualize, providing valuable insights into the decision-making process. Decision trees are widely used for their simplicity, transparency, and effectiveness in handling a variety of predictive tasks.

**Random Forest Algorithm**

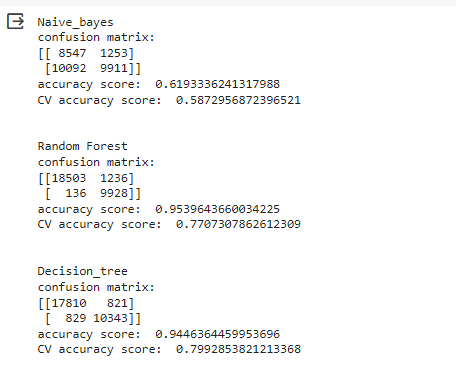
The Random Forest algorithm is a widely used supervised machine learning algorithm capable of handling both classification and regression problems in the field of machine learning. Unlike a single decision tree, the Random Forest algorithm incorporates multiple decision trees for its predictions. Instead of relying on the output of a single tree, the algorithm aggregates predictions from each tree and makes its final prediction based on the majority vote.

In essence, the Random Forest algorithm leverages an ensemble of decision trees, which collectively contribute to the final output. This approach enhances the algorithm's robustness and generalization capabilities. Additionally, the utilization of multiple trees helps mitigate the risk of overfitting, promoting a more accurate and reliable prediction. The Random Forest algorithm is recognized for its versatility, effectiveness, and ability to deliver robust results across various types of machine learning tasks.

**Naive Bayes Algorithm**

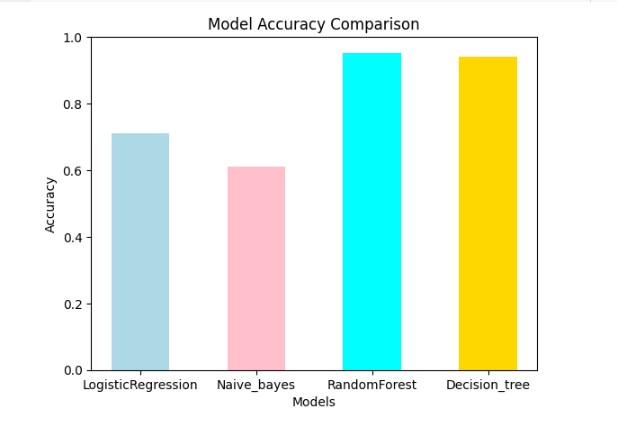
Naive Bayes is a simple yet effective supervised machine learning algorithm commonly used for classification tasks. It operates on the assumption of feature independence, making it computationally efficient. It calculates the probability of a data point belonging to a specific class based on its features and predicts the class with the highest probability. Despite its simplicity, Naive Bayes is often suitable for real-time applications and large datasets, particularly when feature independence is reasonable for the problem at hand.





1. **Model Evolution and comparison**

At this stage, the model undergoes evaluation, and a comprehensive analysis is conducted on the steps taken to ensure that the chosen model achieves optimal quality in addressing the business problem. The evaluation metrics include accuracy, recall, precision, and F1 values, each generated by the respective models. These metrics serve as key indicators to gauge the effectiveness and performance of the models, aligning with the objectives of the business problem at hand.



1. **Conclusion:**

* hotel\_booking.csv dataset is a supervised classification dataset.
* Among these four machine learning algorithms, Random forest and Decision trees perform well with respect to accuracy.
* This models classifies whether or not a booking will be canceled with 95% and 94% accuracy.
* Our EDA (Exploratory data analysis ) explain a non-refundable deposit and requiring parking spaces are the two features influencing cancellation prediction the most.
* Our analysis also pointed at the importance of lead time, number of special requests, and room type.
* we discovered that a deeper understanding of the situation may require additional hotel specific information (such as surrounding parking availability or deposit policies).

1. **Recommendation**

Some recommendations to improve model performance with better accuracy of F1 prediction:

1. Feature Enrichment: Enhance the model by introducing new features or fields related to guest information. Consider incorporating variables like lead time, reservation time histories, and details about guest complaints. These additional features can provide richer insights into booking patterns and cancellation behaviors.

2. Algorithm Exploration: Experiment with alternative machine learning algorithms that may offer improved predictive performance. Regularly explore new algorithms to stay updated with advancements in the field. Additionally, fine-tune hyperparameters of the current model to optimize its performance.

3. Seasonal Adjustments: Account for seasonal variations, especially during holidays like Summer, Christmas, and New Year. Adjusting the model or implementing specialized models to handle holiday-specific trends can enhance prediction accuracy during these periods.

4. Error Analysis: Analyze instances where the model predicts incorrectly to identify patterns and reasons behind these inaccuracies. Understanding the characteristics of mispredictions can guide further refinements to the model, leading to increased accuracy and reliability.

By incorporating these strategies, the model can evolve and adapt to changing conditions, ensuring its effectiveness in predicting hotel booking cancellations.

1. **References**

* <https://www.analyticsvidhya.com/blog/2022/03/end-to-end-hotel-booking-cancellation-machine-learning-model/>
* <https://www.kaggle.com/datasets/hoyeonkangjoe/hotel-booking-cancellation-prediction>
* <https://www.ijraset.com/research-paper/hotel-booking-prediction-using-ml>
* <https://www.sciencedirect.com/science/article/abs/pii/S0167923623000349>
* [https://medium.com/analytics-vidhya/predicting-hotel-booking- cancellations-3fe40d4522f8](https://medium.com/analytics-vidhya/predicting-hotel-booking-%20%20cancellations-3fe40d4522f8)
* <https://www.redalyc.org/pdf/3887/388751309003.pdf>

1. **Bibliography**

1. Original Dataset Paper:

Antonio, N., Almeida, A., & Nunes, L. (2019). Hotel Booking Demand Datasets. Data in Brief, 22, 41-49. https://doi.org/10.1016/j.dib.2018.11.126

2. Machine Learning Algorithms Book:

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

3. Decision Trees and Random Forests Paper:

Breiman, L. (2001). Random Forests. Machine learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

4. Naive Bayes in Text Classification:

Rish, I. (2001). An empirical study of the naive Bayes classifier. IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence, 41-46.