Module 6 – Final Report

ALY 6040: Data Mining Application

Prof. Kasun Samarasinghe

SUBMITTED BY:

Sudeep Ravindra Bedmutha

Trilok Palla

Neelanjan Mitra

Introduction:

The hotel industry is an integral component of the travel and tourism sector, influencing the experiences of travelers and tourists around the world. One of the key challenges faced by hotels is the unpredictability of bookings due to cancellations. Managing and predicting cancellations effectively can lead to better room management, improved customer service, and enhanced profitability.

The dataset in question provides a comprehensive view of hotel bookings, capturing an array of features that detail the booking process and customer preferences. It encompasses variables ranging from basic booking details, such as lead time and type of hotel, to more intricate details like special requests, type of deposit, and even identification of the agent or company involved in the booking.

The data spans multiple years and months, offering insights into patterns over time. Additionally, the presence of columns indicating the number of adults, children, and babies, as well as the type of customer, sheds light on the diverse clientele that hotels cater to. The Average Daily Rate (ADR) and car parking requirements can provide hints into the economic aspects of the bookings.

In this report, we will evaluate the performance of several machine learning models on a dataset and compare their results using various evaluation metrics. The models we will analyze include a Decision Tree Classifier, K-Nearest Neighbors (KNN), Random Forest Classifier, Gradient Boosting Classifier, and XGBoost Classifier. We will assess their performance using metrics like ROC AUC, PR AUC, Lorenz Curve (GINI), and Confusion Matrix. The goal is to identify the best-performing model for the given dataset

Data

The dataset we're working with is systematically organized, consisting of feature variables (X) and a specific target variable (y). These features cover a broad spectrum of data attributes, providing insights into various aspects, while our predictive focus is on the target variable. To guarantee a thorough evaluation of our machine learning models, we've strategically split the dataset into a training subset and a testing subset. The training subset is the cornerstone for building and refining our models. Once these models are established and fine-tuned, they're tested against the testing subset. This method ensures that we can critically assess their efficiency in situations they haven't seen before, making our conclusions about their accuracy both holistic and trustworthy. In essence, this approach underscores the need for an impartial evaluation and guarantees that our models are both versatile and dependable when presented with novel data. Some of the Key Variables used in the Analysis are

hotel: This variable helped differentiate between Resort and City hotels, providing insights into booking behaviors associated with each type.

is_canceled: The primary target variable for our classification models, indicating if a booking was canceled or not.

lead_time: Representing the time between booking and the actual stay, this variable offered insights into booking habits and their potential correlation with cancellations.

arrival_date_month: This allowed for seasonality analysis, indicating booking trends and cancellation patterns across different months.

stays_in_weekend_nights and stays_in_week_nights: These variables provided an understanding of the length of stay and its potential influence on cancellations.

country: Used to visualize the geographical distribution of bookings and potentially understand which regions have higher cancellation rates.

previous_cancellations and previous_bookings_not_canceled: These offered context about the guests' booking history, which can be crucial in predicting future behavior.

booking_changes: Representing the number of changes to a booking, this variable can indicate uncertainty or change of plans, potentially correlating with cancellations.

adr: The Average Daily Rate gave insights into the financial aspects of bookings and their potential correlation with cancellations.

customer_type: Differentiating between types of customers (e.g., transient, groups) can provide insights into different booking and cancellation behaviors.

Exploratory Data Analysis

Use-Case: Predictive Analysis of Booking Cancellations

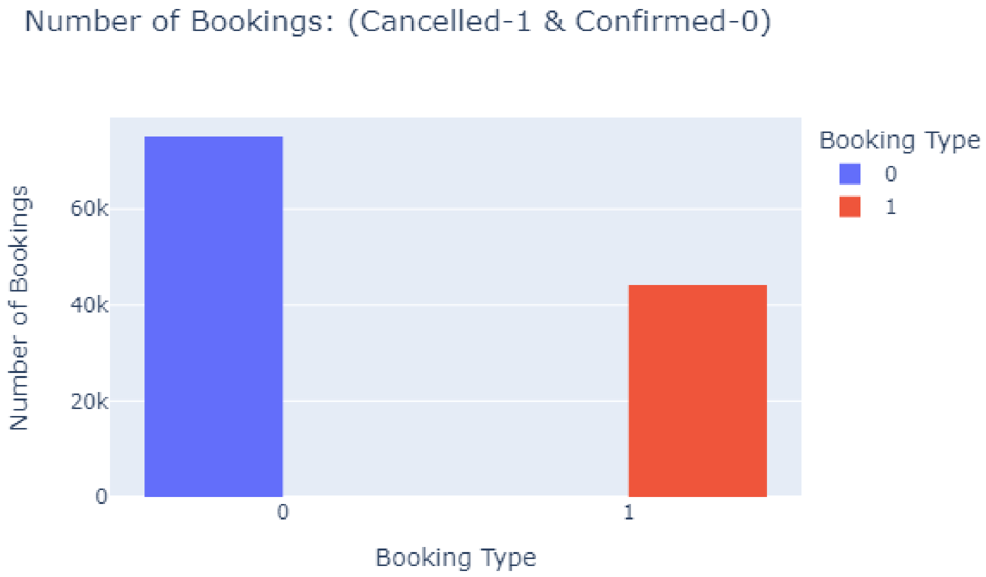
Cancellations are not merely a missed revenue opportunity but also represent an administrative cost and potential disruption in room allocation and services. By leveraging the rich information present in this dataset, we aim to develop a model that can classify customers based on their likelihood of canceling a booking. Such a model can serve as a valuable tool for hotel management, enabling them to:

- Prioritize bookings that have a lower likelihood of cancellation.
- Implement dynamic pricing or offer special deals to customers who are more likely to cancel, to incentivize them to retain their bookings.
- Improve forecasting of room availability and optimize inventory.
- Enhance customer service by understanding the underlying reasons for cancellations and addressing them proactively.
- In the subsequent steps, we will embark on an exploratory analysis of the data, followed by feature engineering and model development to achieve this classification goal.

Here we explored the data to investigate whether there are any patterns in which bookings are being cancelled.

- **Number of Bookings Based on Booking Type**

  Here's the bar plot illustrating the number of bookings based on their type:
  Booking Type 0 (Confirmed): Represents bookings that were not canceled.
  Booking Type 1 (Cancelled): Represents bookings that were canceled.

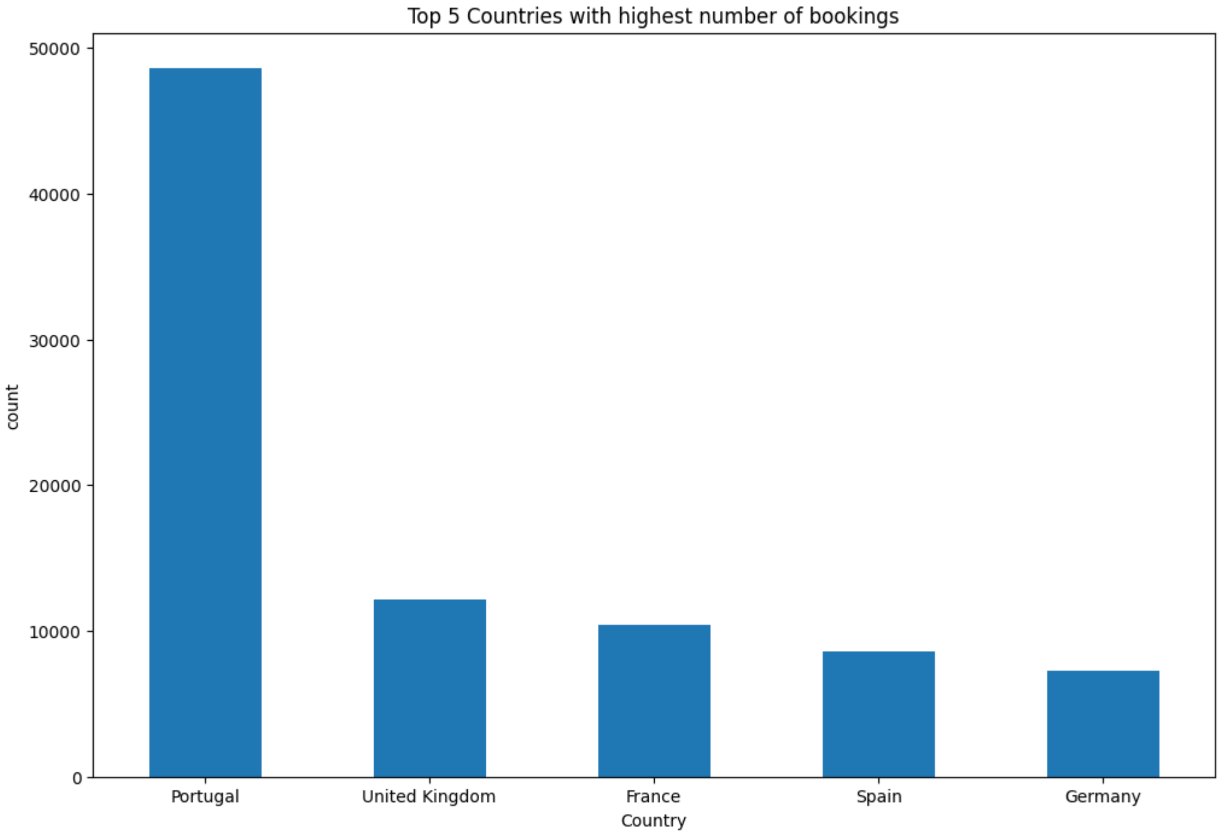


From the graph, we can observe that there are more confirmed bookings (Type 0) than canceled bookings (Type 1), though the number of cancellations is still substantial. This provides insights into the booking behavior of customers and can be a basis for further analysis of factors leading to cancellations.

- **Number Of Records by Countries**

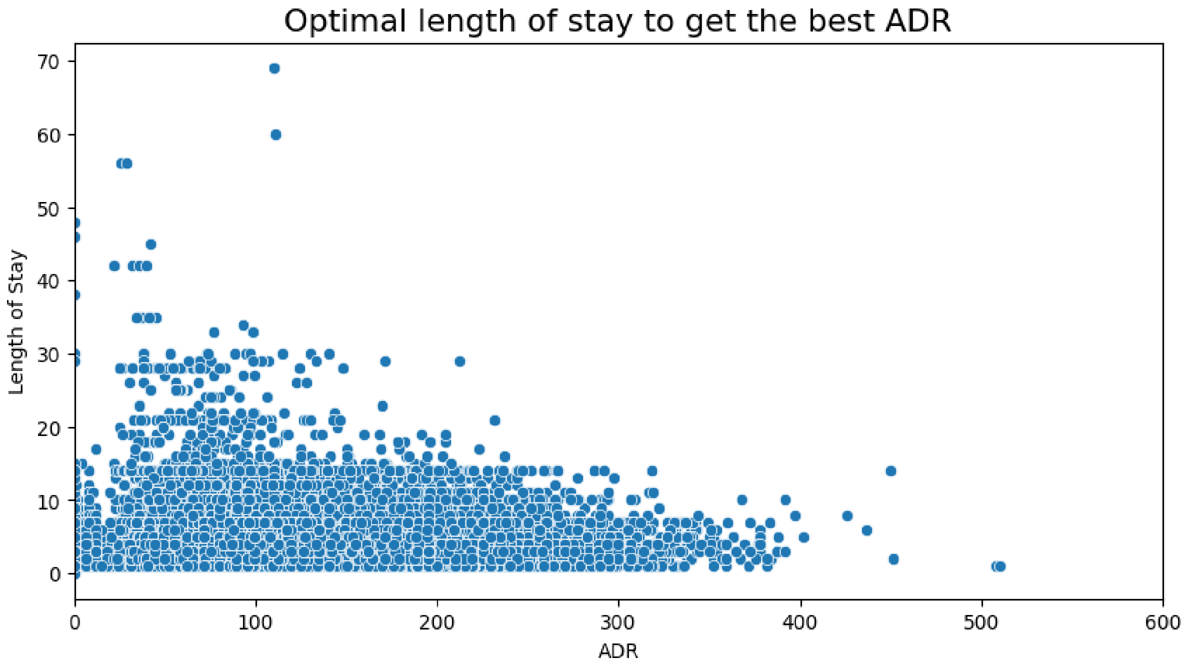
Number of Records by Countries
count
40k
30k
20k
10k
0

Here we have further found the top 5 countries based on the number of records which are,
PRT (Portugal) has the highest number of records with 48,590.
GBR (United Kingdom) follows with 12,129 records.
FRA (France) is next with 10,415 records.
ESP (Spain) has 8,568 records.
DEU (Germany) has 7,287 records.



From the graph, it's evident that Portugal has a significantly higher number of records compared to the other countries in the top 5. This could indicate that the majority of hotel guests or the primary target audience is from Portugal, or the hotel(s) in the dataset are located in or near Portugal.

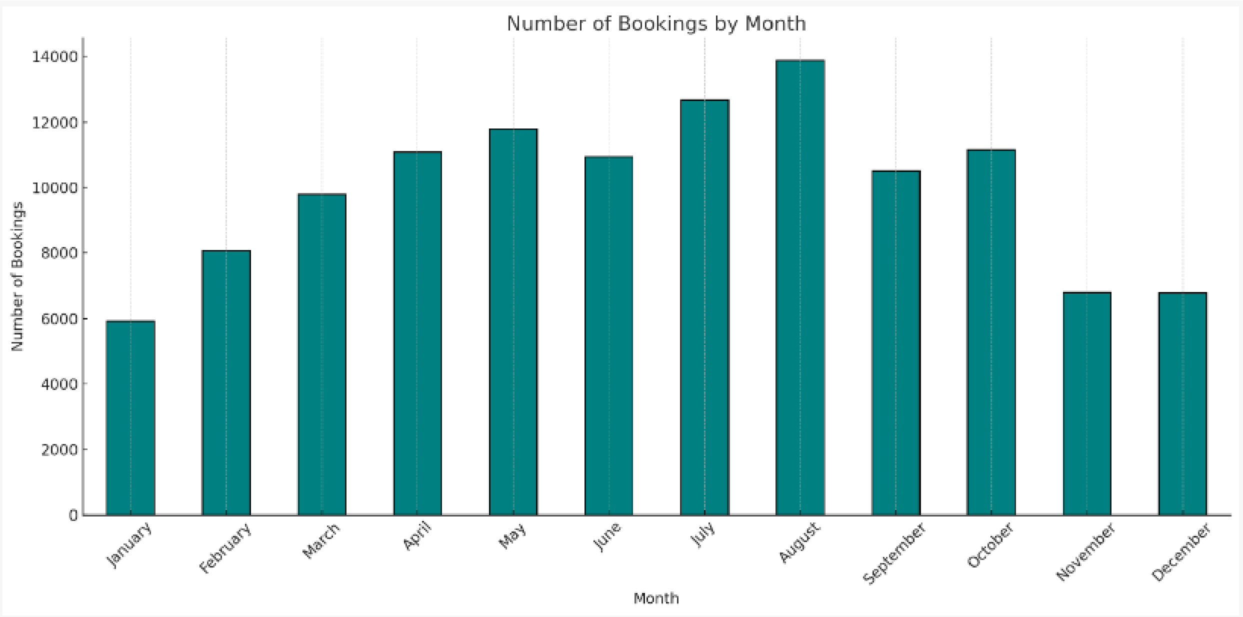
- **Optimal Length of Stay to get the best ADR**



The scatter plot showcasing the relationship between Average Daily Rate (ADR) and Length of Stay reveals diverse guest booking behaviors. A significant cluster of bookings is observed at the lower ADR range, indicating a preference for more affordable rates. While short stays exhibit a wide ADR range, reflecting varied guest preferences from budget to luxury, longer stays show a tighter ADR distribution, suggesting potential price sensitivity for extended durations. Notably, the presence of high ADR values for shorter stays could denote premium bookings, possibly during peak seasons or for luxury suites. The absence of a clear linear trend between ADR and stay length underscores that pricing isn't the sole factor influencing booking duration. Points distant from the main cluster might represent unique packages or outliers. This data suggests that hotels might benefit from a diverse range of offerings to cater to the multifaceted needs and preferences of their guests.

- **Monthly Distribution of Data**

The bar chart illustrates the number of bookings by month:



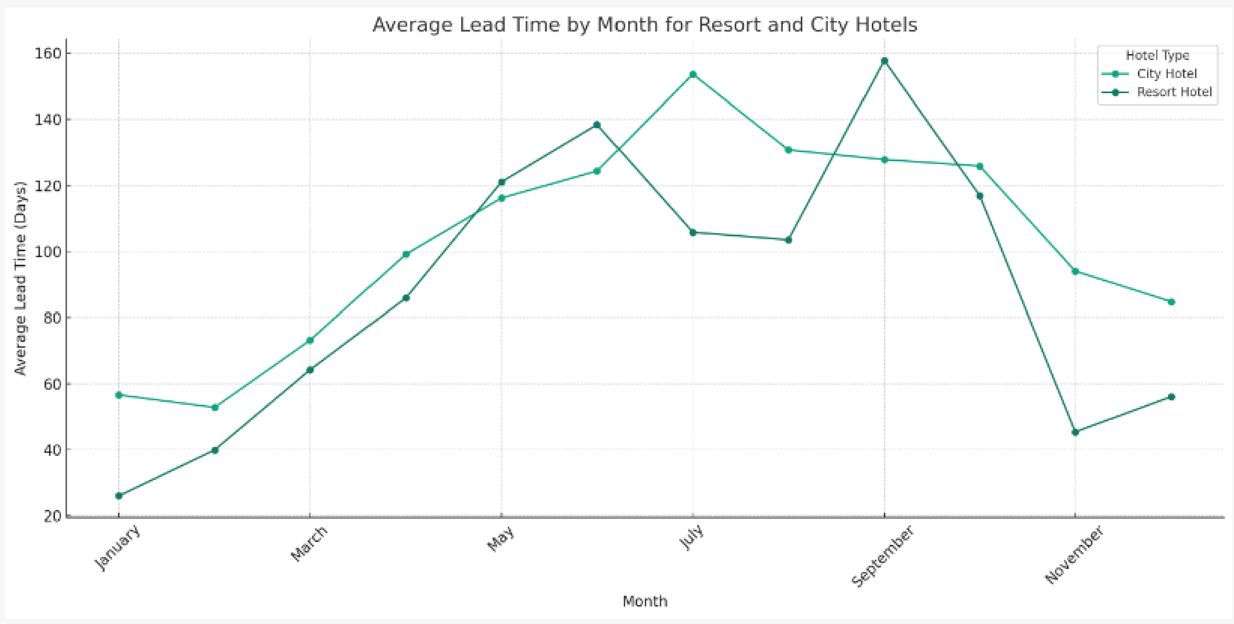
Some of the Insights that we got from this bar chart are as follows,

- July and August seem to have the highest number of bookings, which could be attributed to the summer vacation season in many countries.
- The months following, like September, see a decline, which is expected as the vacation season winds down and schools typically resume.
- The winter months, particularly January and February, appear to have fewer bookings, possibly due to colder weather or fewer vacation days after the holiday season.
- This seasonality insight can be crucial for hotel management to prepare for high-demand months, optimize pricing, and strategize promotions for off-peak periods.

Next, we visualized the lead time for every month, since, from earlier observations, the number of bookings increased during the summer months a similar pattern should be observed here as well.

- **Lead Time Across Months**

The line chart illustrates the average lead time by month for both Resort and City Hotels:



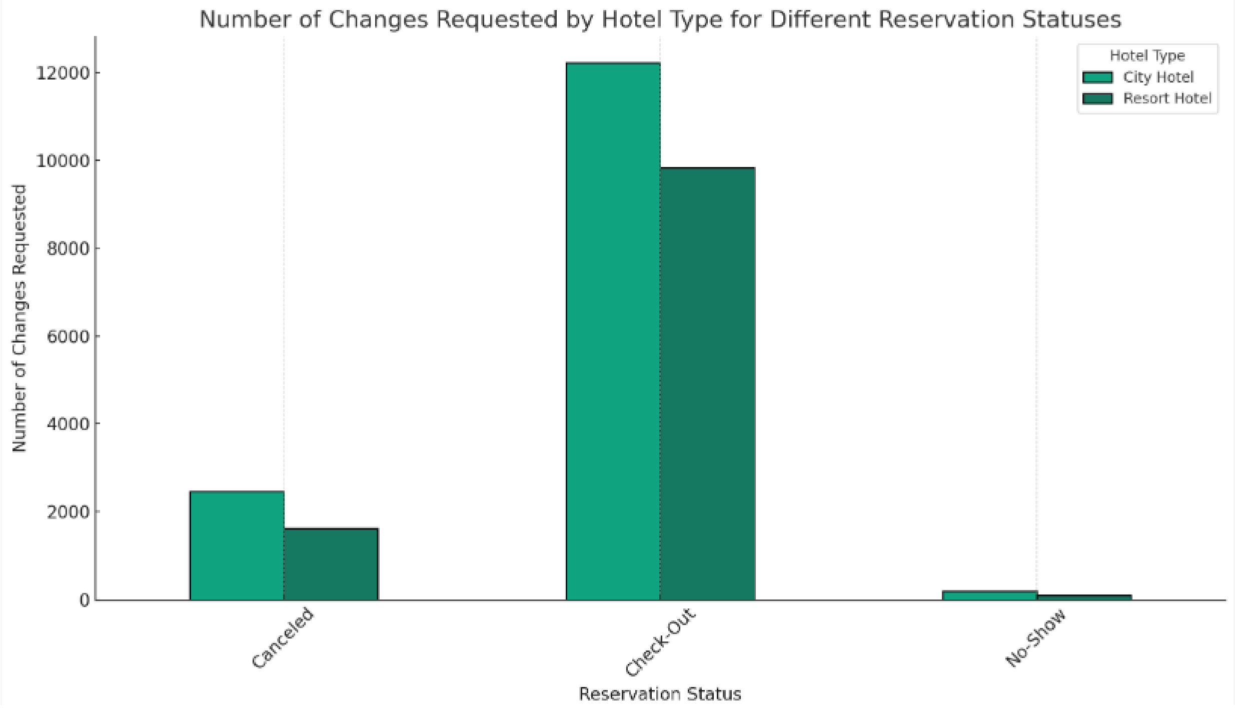
Resort Hotel Trend: The lead time for Resort Hotels peaks during the summer months, particularly in July and August. This indicates that guests tend to book their resort vacations well in advance, especially for the summer season.

City Hotel Trend: City Hotels also see an increase in lead time during the summer months, but it's less pronounced compared to Resort Hotels. The lead time for City Hotels seems to be more evenly distributed throughout the year, suggesting that business or short-term travel might be a significant component of their bookings.

Comparison: Overall, the lead time for Resort Hotels tends to be longer than that for City Hotels, especially in peak vacation months.

This differentiation in booking behaviors between Resort and City Hotels can help hoteliers tailor their marketing strategies, promotions, and pricing models to better cater to their target audiences. For instance, Resort Hotels might focus on early booking promotions for the summer, while City Hotels could offer flexible deals year-round.

- **Reservation Status For Number of Changes Requested**



Here's the bar graph illustrating the number of changes requested based on hotel type (Resort and City Hotels) for different reservation statuses, with booking changes on the Y-axis:

Check-Out: Both City and Resort Hotels have significant changes requested for bookings that eventually check out. City Hotels have a notably higher number compared to Resort Hotels.

Canceled: For bookings that get canceled, City Hotels again witness a higher number of changes compared to Resort Hotels. This trend suggests that before canceling, many guests, especially in City Hotels, try to modify their bookings.

No-Show: For the no-show category, the difference between City and Resort Hotels in terms of changes requested is less pronounced, but City Hotels still have a slightly higher count.

This visualization reaffirms the earlier observation that City Hotels experience more frequent booking modifications than Resort Hotels. The data can provide insights to hotel management on how to handle booking changes efficiently, especially in City Hotels where changes seem more prevalent.

Here are some of the Insights we can make according to our EDA

Booking Distribution:
The majority of bookings were confirmed, but a significant number of reservations were also canceled, indicating potential lost revenue opportunities for the hotels.
Country Origin: Most bookings originated from countries like Portugal, the UK, France, Spain, and Germany. Understanding these primary sources of guests can help in tailoring marketing strategies and services to cater to these nationalities.

Seasonality:
The months of July and August witnessed the highest number of bookings, reflecting peak vacation season trends.
Winter months like January and February had fewer bookings, indicating possible off-peak periods.

Lead Time Trends:
The lead time (time between booking and stay) varied across months, with longer lead times observed during the summer.
City Hotels generally had more frequent booking modifications than Resort Hotels.

Booking Behavior by Hotel Type:
City Hotels experienced more bookings and booking modifications compared to Resort Hotels.
The type of hotel (City vs. Resort) influenced the booking lead time, with City Hotels generally having more short-term bookings.

Booking Modifications:
Most modifications occurred in bookings that were eventually confirmed and checked out, followed by those that were canceled.
Guests who didn't show up for their reservations (No-Shows) were less likely to have made previous changes to their bookings.

Length of Stay:
The length of stay didn't show a clear linear relationship with the Average Daily Rate (ADR). This suggests that other factors, possibly hotel amenities, seasonality, or guest demographics, might influence the duration of stays more than just the price.

Average Daily Rate (ADR):
There was a diverse range of ADR values for different lengths of stay, with certain high ADR values associated with short stays, possibly indicating premium bookings.

Booking Changes by Reservation Status:
Booking modifications were most frequent in confirmed reservations, followed by canceled bookings. No-shows had the least number of changes, indicating that these guests rarely interacted with their bookings before not turning up.

In summary, the EDA provided valuable insights into guest booking behaviors, seasonality trends, and preferences based on hotel type. These insights can be instrumental for hotel management to optimize pricing, improve guest services, tailor marketing strategies, and enhance overall guest experience.

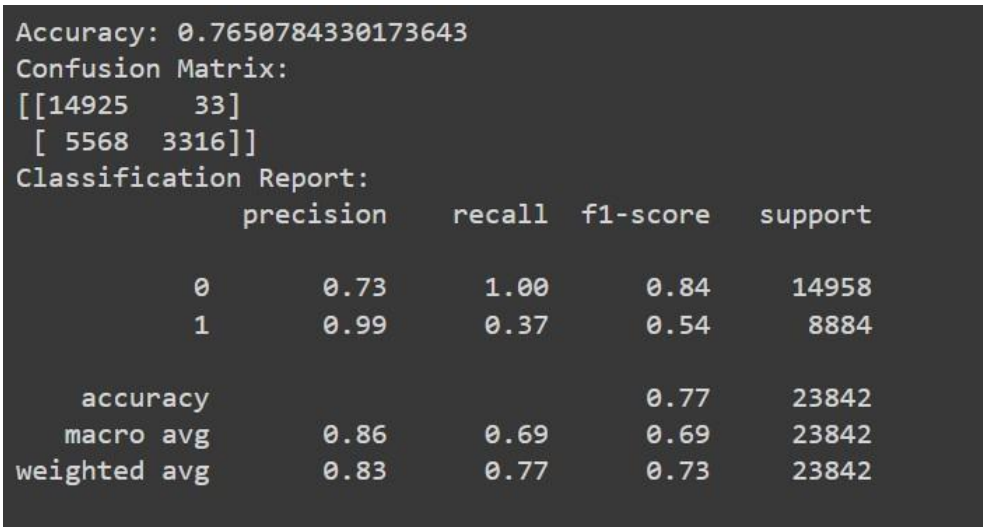
Models Used for Evaluating Performance

Decision Tree Model:

We start by evaluating a Decision Tree Classifier on the dataset. We also use hyperparameter tuning to find the best combination of hyperparameters for the Decision Tree model. The following results were obtained:

Decision Tree Model Evaluation:

Initial Accuracy: 0.7650



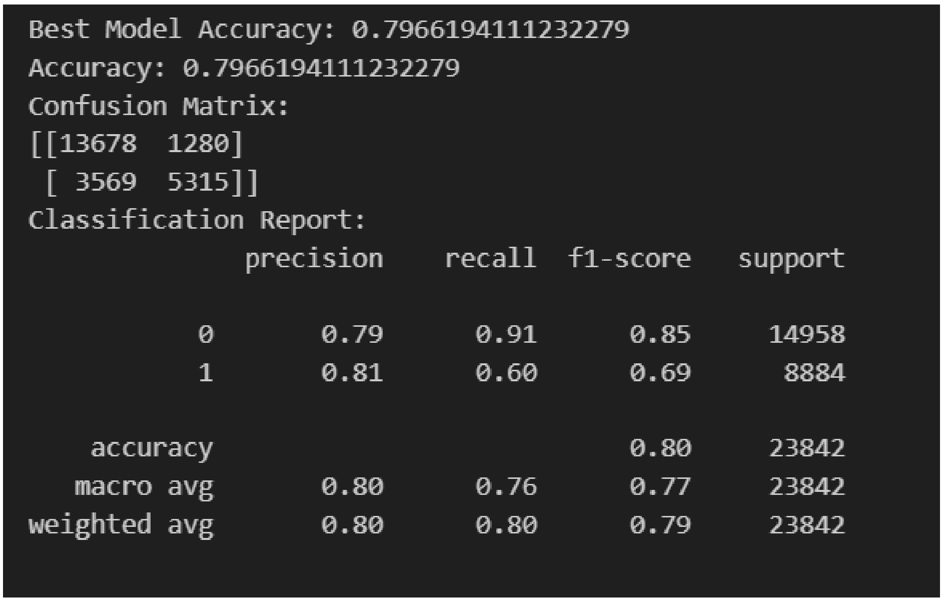
```
Accuracy: 0.7650784330173643
Confusion Matrix:
[[14925    33]
 [ 5568  3316]]
Classification Report:
              precision    recall  f1-score   support

           0       0.73      1.00      0.84     14958
           1       0.99      0.37      0.54      8884

    accuracy                           0.77     23842
   macro avg       0.86      0.69      0.69     23842
weighted avg       0.83      0.77      0.73     23842
```

Figure 1: Decision Tree

- The model exhibits a high precision for class 1 but a low recall. This means while it's very confident when it predicts class 1, it misses out on a significant number of actual class 1 instances.
- The model seems to favor class 0 predictions, as evidenced by the high recall for class 0 and the high false negatives for class 1.
- Depending on the application, the trade-off between precision and recall might be acceptable. For instance, if the cost of false positives is high, a higher precision is desirable. However, if missing out on actual positives (false negatives) is costly, the model might need further tuning or improvement to increase recall for class 1.
- It might be beneficial to explore techniques like resampling, using different algorithms, or feature engineering to improve the model's performance, especially for class 1 recall.

Decision Tree Classifier:

Final Accuracy: 0.7966



```
Best Model Accuracy: 0.7966194111232279
Accuracy: 0.7966194111232279
Confusion Matrix:
[[13678  1280]
 [ 3569  5315]]
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.91      0.85     14958
           1       0.81      0.60      0.69      8884

    accuracy                           0.80     23842
   macro avg       0.80      0.76      0.77     23842
weighted avg       0.80      0.80      0.79     23842
```

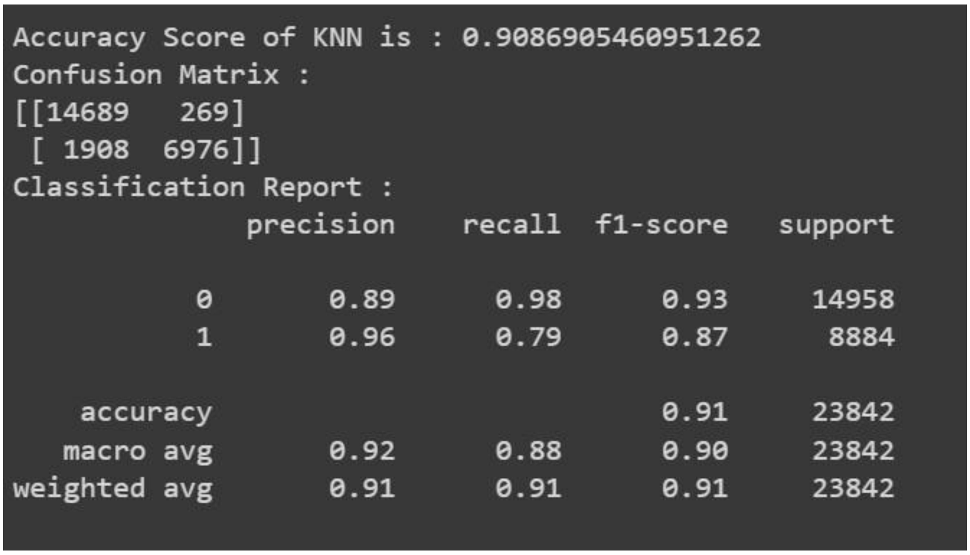
- The Decision Tree model with optimized hyperparameters shows an improved performance compared to the previous model, especially in terms of accuracy and recall for class 1.
- While the model has a high precision for both classes, the recall for class 1 is still a bit on the lower side, indicating the model misses out on a significant number of actual class 1 instances.
- The improved F1 scores suggest that the balance between precision and recall is better with the optimized hyperparameters.
- Despite the enhancements, further tuning or even trying different algorithms might help in achieving even better results, especially in increasing the recall for class 1 without compromising the precision.

KNN Model:

Next, we assess the performance of a K-Nearest Neighbors Classifier on the dataset. The following results were obtained:

Evaluation:

Accuracy: 0.9087



```
Accuracy Score of KNN is : 0.9086905460951262
Confusion Matrix :
[[14689   269]
 [ 1908  6976]]
Classification Report :
              precision    recall  f1-score   support

           0       0.89      0.98      0.93     14958
           1       0.96      0.79      0.87      8884

    accuracy                           0.91     23842
   macro avg       0.92      0.88      0.90     23842
weighted avg       0.91      0.91      0.91     23842
```

The model's accuracy is an impressive 90.87%90.87%. This indicates that the model correctly predicted the outcome for about 90.87%90.87% of the samples in the test set. This is higher than both the previous models we discussed.

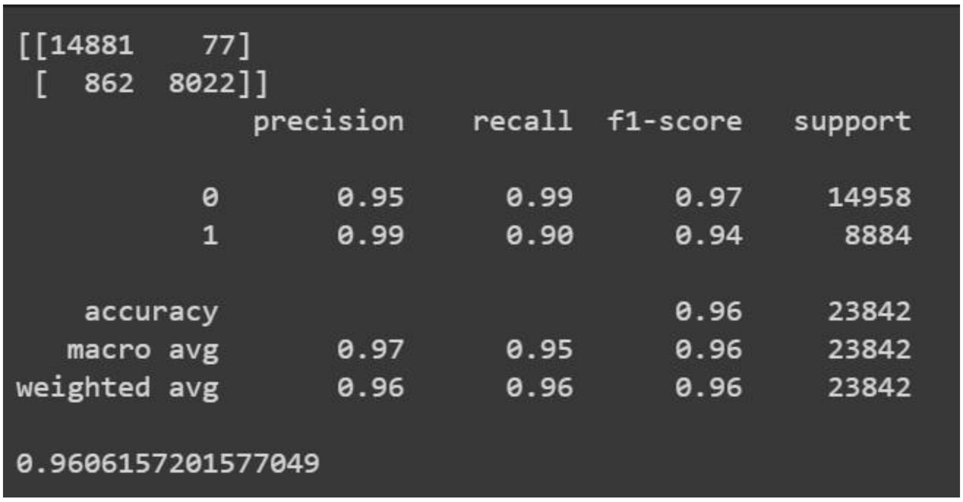
- The KNN model shows a very strong performance across all metrics when compared to the previous models. It has high accuracy, precision, and recall.
- The model's high precision for class 1 suggests that when it predicts an instance as class 1, it's highly likely to be correct.
- The recall for class 1 has also improved significantly, meaning the model is capturing a greater proportion of actual class 1 instances.
- The F1 scores further validate that the KNN model achieves a good balance between precision and recall, making it a robust choice for this dataset.

Random Forest Classifier:

We then examine the performance of a Random Forest Classifier. The following results were obtained:

Evaluation:

Accuracy: 0.9606



```
[[14881    77]
 [  862  8022]]
              precision    recall  f1-score   support

           0       0.95      0.99      0.97     14958
           1       0.99      0.90      0.94      8884

    accuracy                           0.96     23842
   macro avg       0.97      0.95      0.96     23842
weighted avg       0.96      0.96      0.96     23842

0.9606157201577049
```

- The Random Forest Classifier demonstrates exceptional performance across all metrics. Its ensemble nature, where it aggregates results from multiple decision trees, contributes to its robustness and ability to handle complex datasets effectively.
- The model's high precision, especially for class 1, indicates that when it labels an instance as class 1, it's very likely to be correct. This can be crucial if the cost of false positives is high.

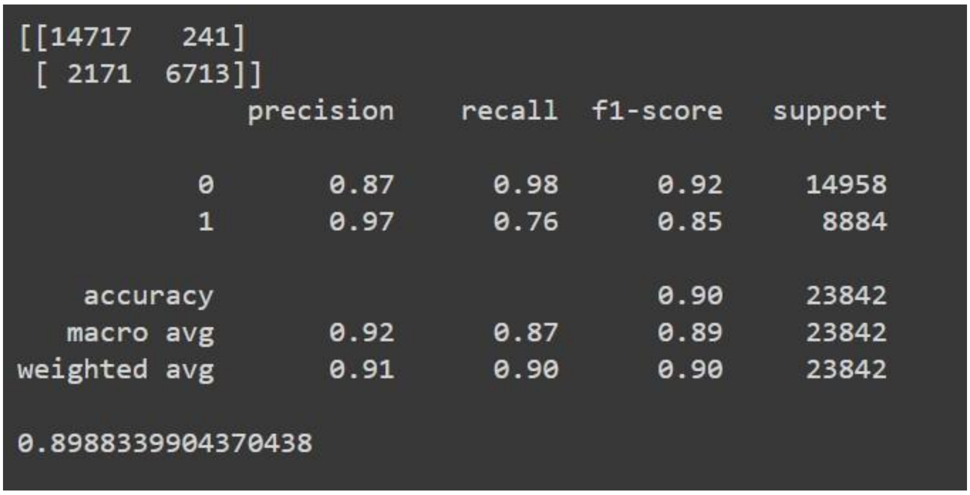
Since Random Forest has Already very High Scores we are not tuning the Model

Gradient Boosting Classifier:

Next, we evaluate the performance of a Gradient Boosting Classifier. The following results were obtained:

Evaluation:

Accuracy: 0.8988



```
[[14717   241]
 [ 2171  6713]]
              precision    recall  f1-score   support

           0       0.87      0.98      0.92     14958
           1       0.97      0.76      0.85      8884

    accuracy                           0.90     23842
   macro avg       0.92      0.87      0.89     23842
weighted avg       0.91      0.90      0.90     23842

0.8988339904370438
```

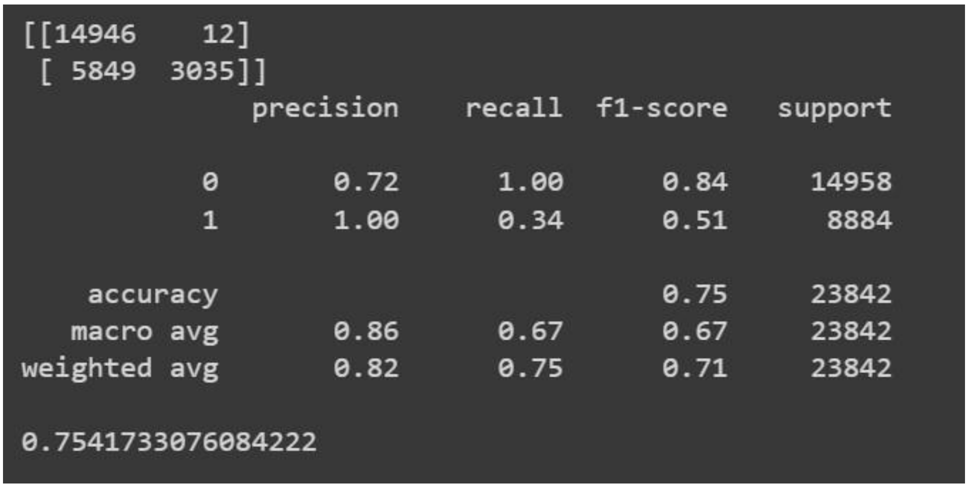
Figure 4: Gradient Boosting Model

XG Boost Classifier:

Finally, we assess the performance of an XGBoost Classifier. The following results were obtained:

Evaluation:

Accuracy: 0.88



```
[[14946    12]
 [ 5849  3035]]
              precision    recall  f1-score   support

           0       0.72      1.00      0.84     14958
           1       1.00      0.34      0.51      8884

    accuracy                           0.75     23842
   macro avg       0.86      0.67      0.67     23842
weighted avg       0.82      0.75      0.71     23842

0.7541733076084222
```

Figure 5: XG Boost Model

Model Comparisons:

While Accuracy is an important metric, it's also crucial to consider other performance metrics and perform additional analysis to ensure that the model's performance is acceptable for the specific problem you are trying to solve. Depending on the nature of your problem, you may need to balance accuracy with other metrics like precision, recall, F1-score, or area under the ROC curve (AUC) to evaluate the model more comprehensively. We compared the performance of the models using various metrics. The ROC AUC scores and PR AUC scores for each model are as follows:

ROC- Area under the Curve scores:

- Decision Tree: 0.82
- Random Forest: 1.00
- Gradient Boosting: 0.95
- KNN: 0.96
- XGBoost: 0.88

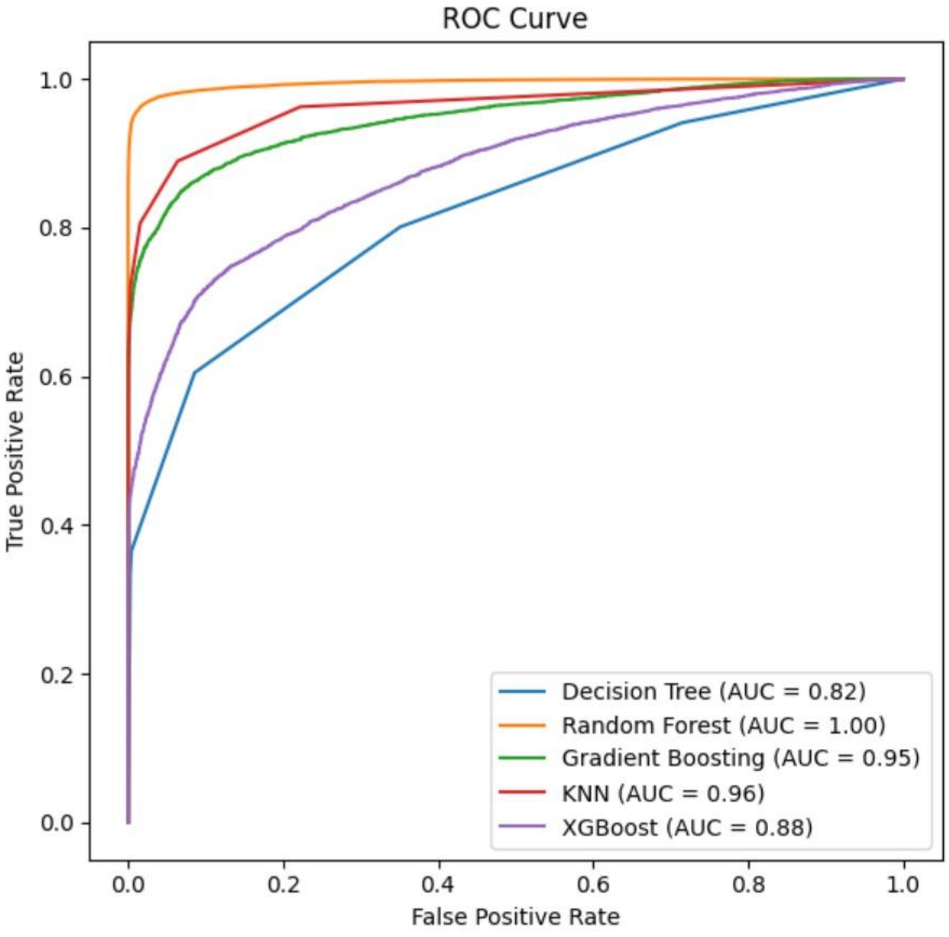
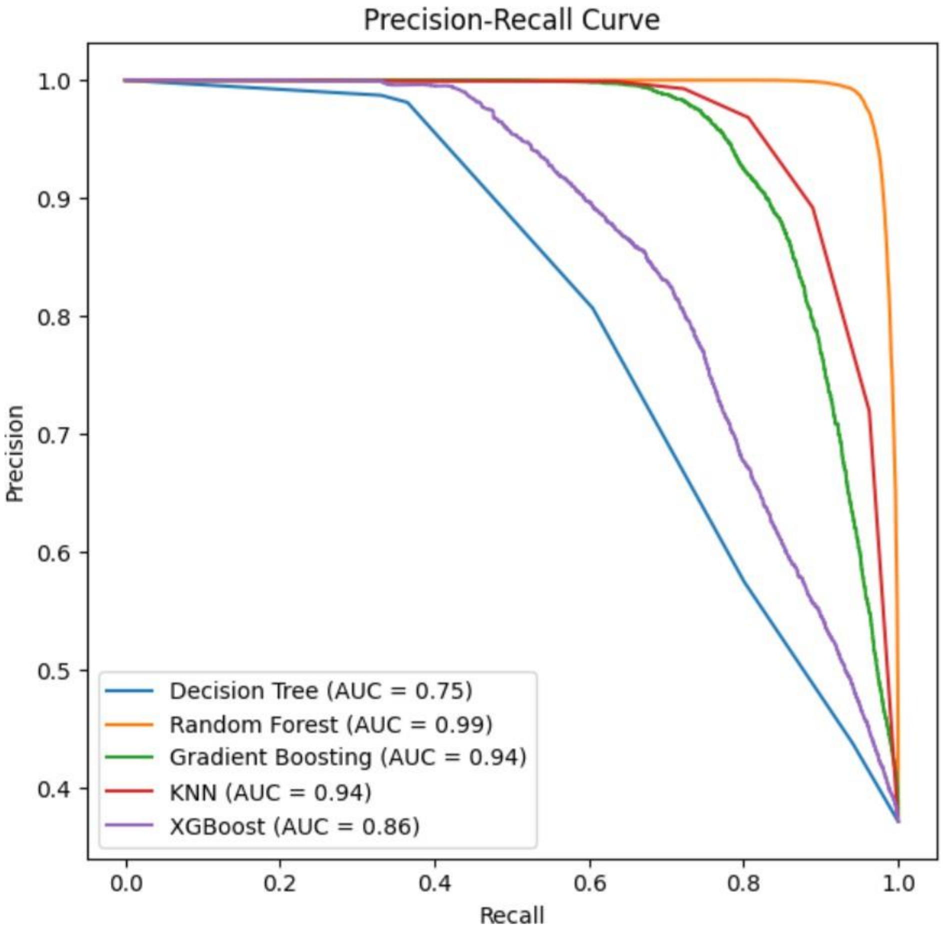


Figure 6: ROC

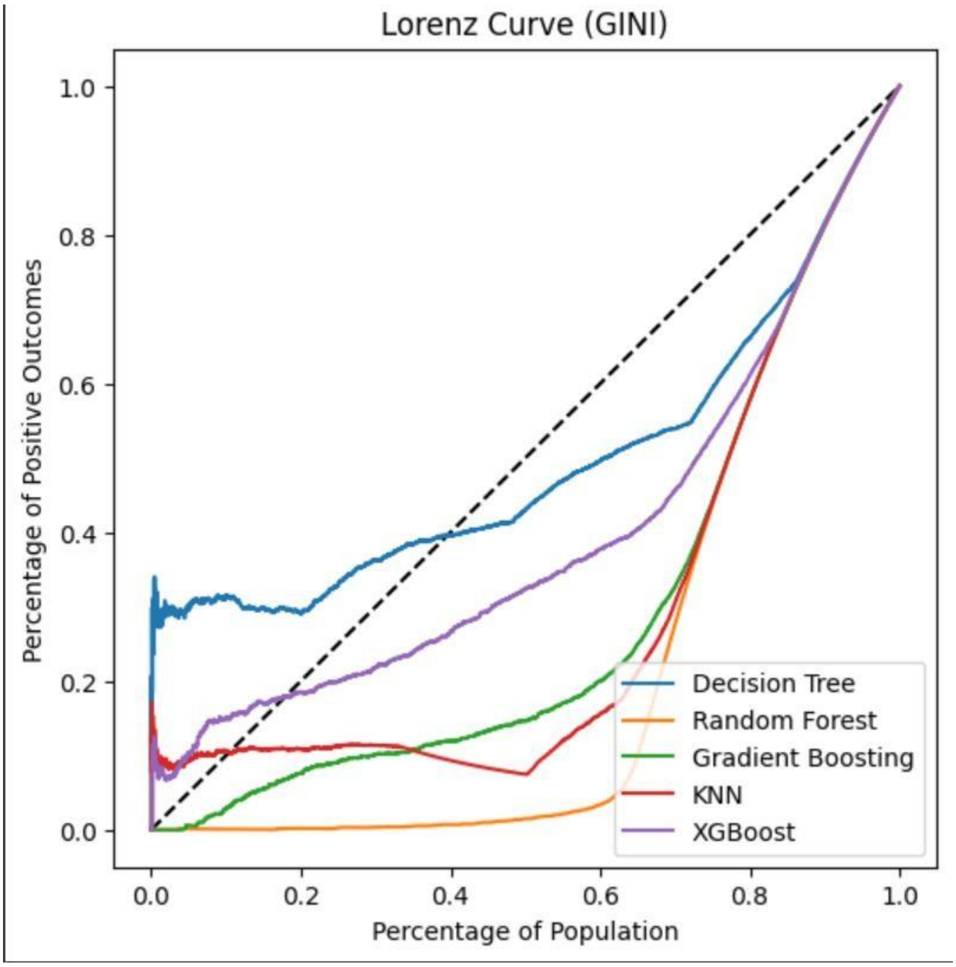
PR Area Under the Curve scores:

- Decision Tree: 0.75
- Random Forest: 0.99
- Gradient Boosting: 0.94
- KNN: 0.94
- XGBoost: 0.86



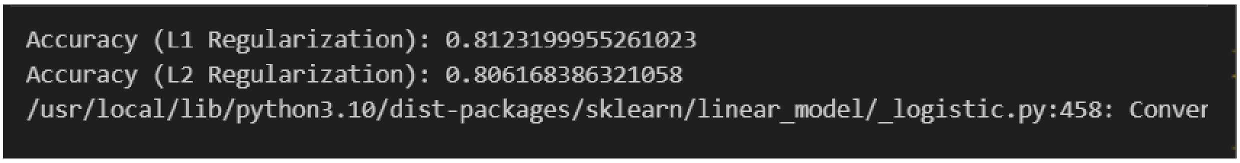
Lorentz Curve:

We also plotted Lorenz curves to visualize the models' performance in terms of the GINI coefficient. The curves indicate the percentage of positive outcomes as a function of the percentage of the population. Higher GINI values imply better performance.



Regularization :

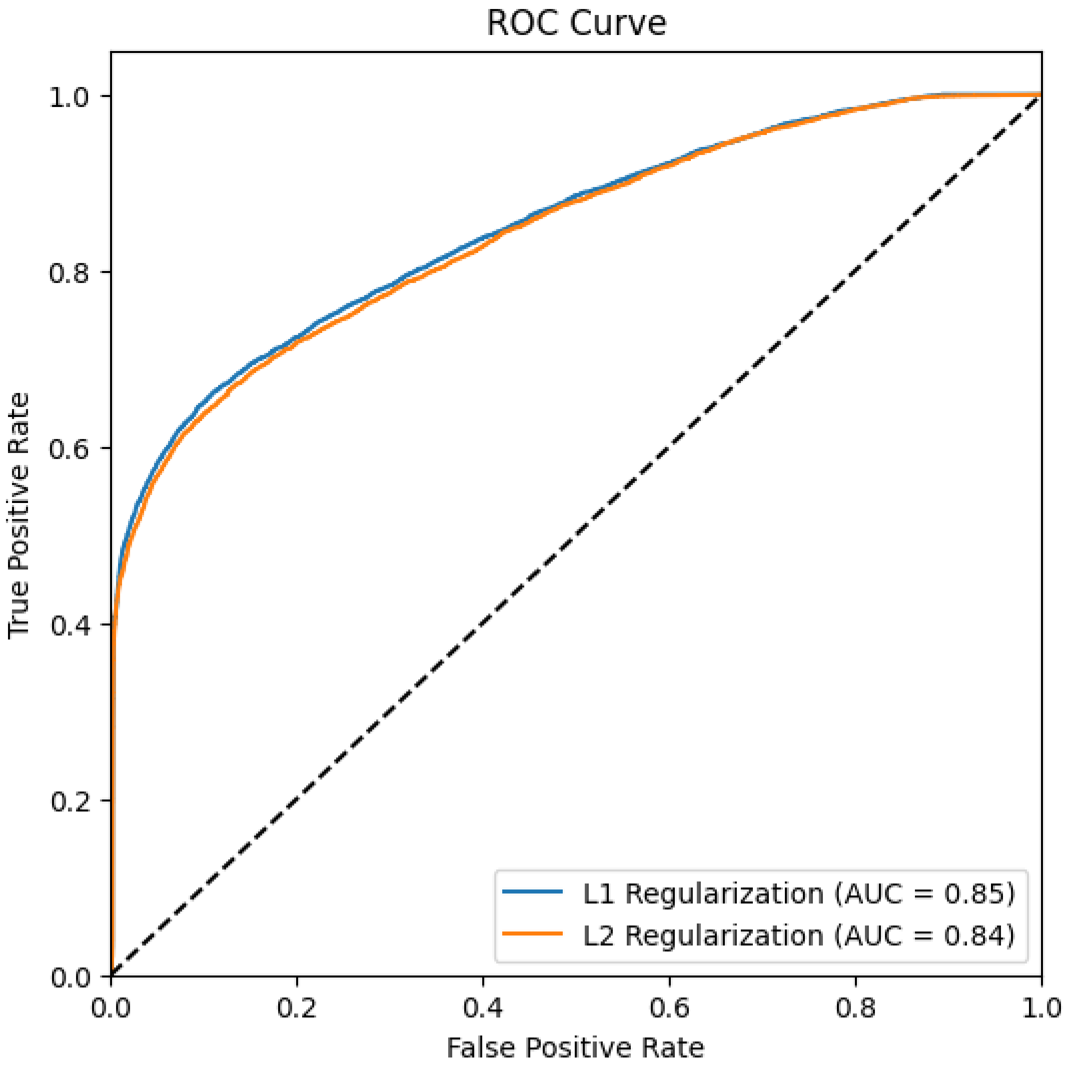
We created, trained, and evaluated two logistic regression models with different types of regularization: L1 (Lasso) and L2 (Ridge) regularization. Here is the Accuracy we have obtained

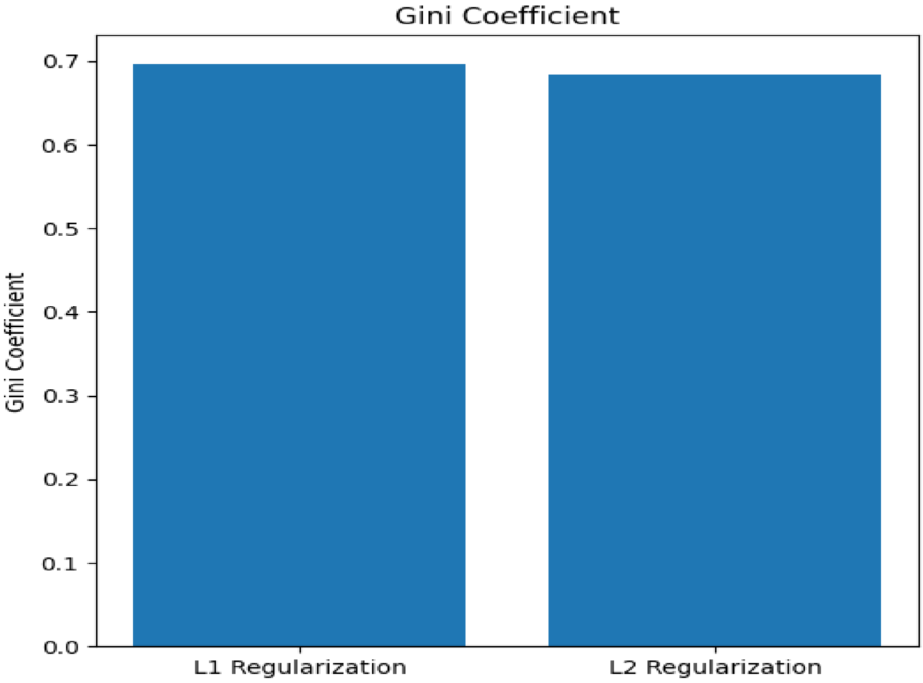


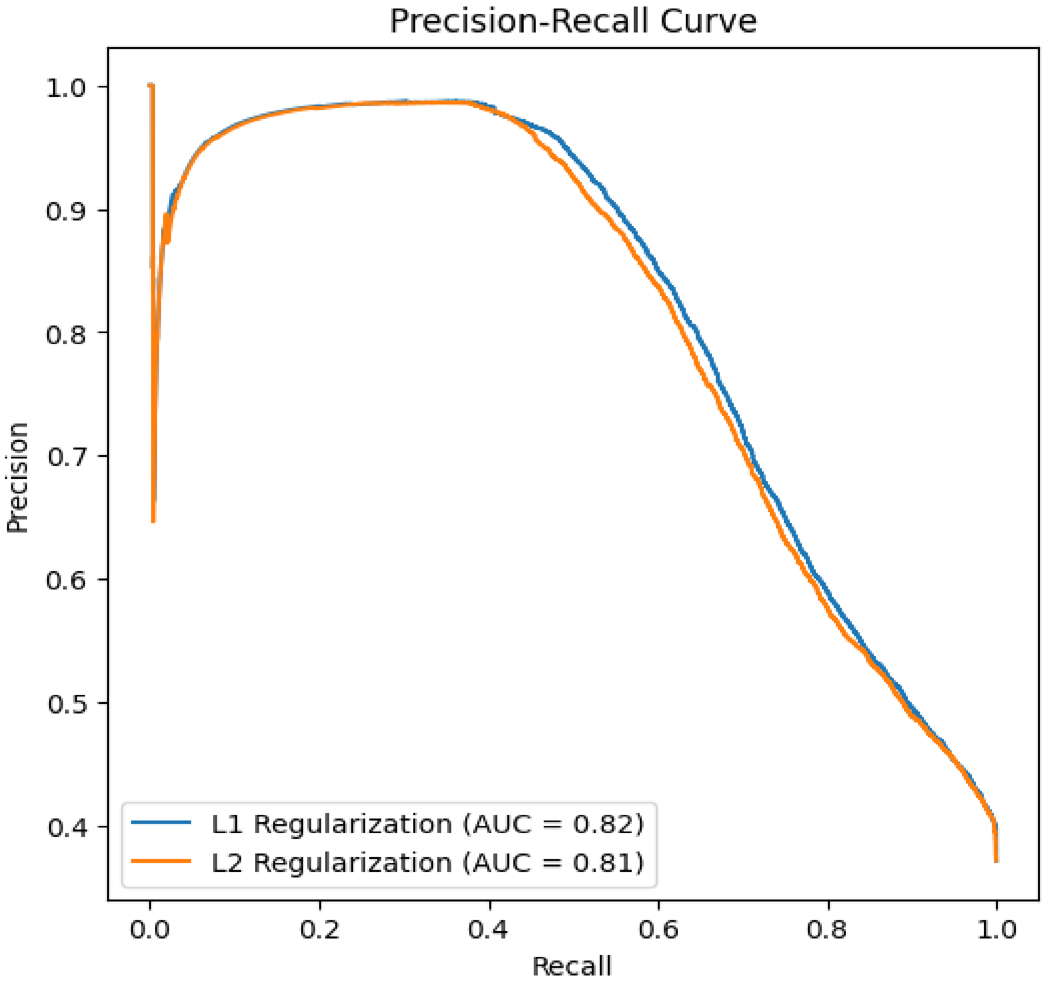
```
Accuracy (L1 Regularization): 0.8123199955261023
Accuracy (L2 Regularization): 0.806168386321058
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458: Conver
```

- The model with L1 regularization performed slightly better than the one with L2 regularization for this specific dataset.
- The difference in performance, while not huge, might be attributed to the feature selection property of L1 regularization. It's possible that some features in the dataset don't contribute much to the prediction, and L1 regularization helped in disregarding them.
- The choice between L1 and L2 regularization often depends on domain knowledge, the nature of the data, and the problem at hand. In this instance, L1 seems to be the more suitable choice.

Precision recall and ROC curve for Logistic Regression models for L1(Lasso) & L2(Ridge) regularizations




Gini Coefficient
Gini Coefficient
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
L1 Regularization
L2 Regularization


Precision-Recall Curve
Precision
1.0
0.9
0.8
0.7
0.6
0.5
0.4
Recall
0.0
0.2
0.4
0.6
0.8
1.0
L1 Regularization (AUC = 0.82)
L2 Regularization (AUC = 0.81)

Recommendations:

Model Selection:
Given the superior performance of the Random Forest Classifier and Gradient Boosting, it's advisable to consider these deployment models, especially if the primary metric of interest is accuracy.

Feature Importance:
With tree-based models like Random Forest and Gradient Boosting showcasing promising results, it's recommended to evaluate feature importance. This can provide insights into which features are most influential in driving predictions, allowing for more informed decision-making and potential feature engineering.

Model Tuning and Validation:
Continue to explore hyperparameter tuning, especially for models like XGBoost which have a plethora of parameters. This can further optimize model performance.
Implement cross-validation techniques to ensure that the models are robust and not just performing well on a specific split of the data.

Addressing Model Complexity:
While ensemble methods have shown strong performance, they can be computationally intensive. If real-time predictions or scalability becomes a concern, consider simplifying the model or exploring more lightweight algorithms.

Evaluation Beyond Accuracy:
Although accuracy is a valuable metric, it's essential to evaluate models based on other metrics like precision, recall, and the F1-score, especially when the costs associated with false positives and false negatives differ.

Consider the business implications: if wrongly predicting a cancellation has a significant business cost, prioritize precision or other relevant metrics.

Feature Reduction with L1 Regularization:
The slight performance edge of Logistic Regression with L1 regularization suggests that there might be redundant or non-informative features. Consider using L1 regularization for feature selection, which can lead to a simpler, more interpretable model.

Model Interpretability:
For critical business decisions, model interpretability can be crucial. It might be worth exploring models or tools that provide insights into how decisions are made, ensuring stakeholders understand and trust the model's predictions.

Conclusion:

Based on the evaluation and comparison of the models, the Random Forest Classifier emerged as the standout, achieving the highest accuracy and PR AUC score. K-Nearest Neighbors and Gradient Boosting also demonstrated commendable performance, proving their utility for such types of datasets. In contrast, the XGBoost model lagged with notably lower accuracy and PR AUC score, suggesting it might not be the best fit for this specific dataset. It's crucial to remember that the efficacy of a model can often depend on the nature of the data and the specific problem context. Therefore, while XGBoost has its strengths in many scenarios, for this particular dataset and problem statement, other models proved to be more adept. .If we want a good balance between precision and recall and value both equally, KNN might be an ideal choice. It has high accuracy and significantly improved recall for class
If we prefer a simpler model and are willing to trade off some accuracy for interpretability, Decision Tree Classifier with optimized hyperparameters is an option.
The Gradient Boosting Classifier and XG Boost Classifier are also reasonable options, but their accuracy is slightly lower than Random Forest and KNN.Also Future endeavors could explore feature engineering, different preprocessing techniques, or even ensemble methods to enhance model performance further.

REFERENCE:

[1] Ian Dickerson and Tom Button (2017). “Hotel Bookings”. Retrieved from https://www.kaggle.com/datasets/mathsian/hotel-bookings/data

[2] https://www.researchgate.net/publication/329286343_Hotel_booking_demand_datasets