MIS710 – Machine Learning in Business - Trimester 1 2023

Assessment Task 1 – Case Study (Report) – Individual

STUDENT ID:
222294384

STUDENT NAME:
ABIADE BISI-KAZEEM

Table of Contents

Executive Summary **3**
Business Understandings & Problem Statement **3**
Data Understandings, Data Preparation, Exploration, and Visualisation **4**
About Data: 4
Data Cleaning 5
Machine Learning Approach **5**
Model and Performance Metrics **5**
Pros & Cons of the model **5**
Business Solution & Recommendations **5**
References **5**

Executive Summary

The impact of cancellations on the hotel business can be significant. Last-minute cancellations can result in lost revenue, wasted resources, and disappointed customers. Vera Selection Resort (VSR), an hospitality business that owns several companies, is looking to expand its business in Melbourne, Victoria. As part of their expansion plans, they want to improve their performance and reduce room cancellations. To achieve this, VSR has provided a dataset with labels generated from their booking and reception systems. This dataset contains information about customer bookings and whether they arrived or canceled their bookings.

Our analysis of the dataset revealed that the most significant factors affecting room cancellations were the type of booking, number of special requests, number of adults, month of booking, lead time, and airport pickup. We developed a machine learning model that can predict the likelihood of a room cancellation, achieving an AUC score of 0.95 and an accuracy of 89% on the test set.

To develop the model, we used a pipeline that includes preprocessing steps such as one-hot encoding, label encoding, and standard scaling. The final model used XGBoost, a powerful machine learning algorithm that is well suited for classification problems.

Overall, our analysis provides valuable insights for VSR and can help them reduce room cancellations by targeting factors such as the type of booking, booking lead time, and the month when bookings are made.

Business Understandings & Problem Statement

Business Understanding

Vera Selection is a business group that operates several businesses in Victoria, Melbourne, including Vera Selection Resort (VSR). VSR is a popular holiday destination and offers luxurious accommodations, recreational activities, and fine dining to its guests. However, the resort has been experiencing a high number of room cancellations, which is negatively impacting their revenue. VSR wants to explore their business data to understand their business better, improve performance, and reduce room cancellations.

Business Problem

One of the main business problems that VSR is facing is a high number of room cancellations, which is a common challenge for many hotels and resorts. When guests cancel their bookings, the resort loses revenue and may have to incur additional costs to fill the empty rooms. Moreover, last-minute cancellations can make it challenging for the resort to manage their resources effectively, which can impact the overall guest experience. In the case of VSR, the management team has noticed that the number of cancellations is higher than they would like, and they want to explore ways to reduce them.

Aim of the Project

The aim of this project is to explore the dataset provided by VSR and develop a machine learning model that can predict room cancellations based on the booking details. The project will identify the factors that are contributing to the high number of cancellations and provide insights and recommendations to VSR. To achieve this aim, the project will involve several stages, including data exploration, data preprocessing, feature engineering, model selection, and model evaluation.

Data Understandings, Data Preparation, Exploration, and Visualisation

About Data

Feature	Description
Booking_ID	Booking ID
RoomType	Room types: Neptune, Venus, Jupiter, Saturn, Mercury, Mars
FromDate	The arrival date of the booking
Adults	Number of adults
Children	Number of children
Breakfast	Types of breakfast services: Buffet Breakfast, Luxury Buffet Breakfast, Super VIP Breakfast, Not Selected
Number of Special Requests	Number of special requests placed by the customer
Airport Pickup	0 means Not required and 1 means Airport pickup required
Weekend Nights	Number of weekend nights booked by the customer

Feature	Description
Weeknights	Number of weeknights booked by the customer
Loyalty Points	Total loyalty points accumulated
Lead time in days	The number of calendar days between the date the booking is made and the actual arrival date
Booking Type	Types of bookings: Online, Offline, Corporate, Complementary, and Aviation
Existing Customer	0 means Not existing and 1 means Existing customer
Previous Cancellations	Number of previous bookings that were canceled by the customer before the current booking
Average Room Price	Average price per day of the booking in AUD
Booking Status	Arrived: the customer has arrived; Canceled: the booking was canceled.

The data used was provided by VSR and it originally contained 32,647 records and 17 features. These features and their description are in the table above.

Data Cleaning

The data contained no duplicates and no null values but there were some inconsistencies and data integrity issues. For example, the "FromDate" feature had values like "FromDate" and "29/2/2018". The rows with "FromDate" as values were removed and the rows with "29/2/2018" were changed to "28/2/2018". The data type for all features were validated and changed to correct types. Removing some records due to invalid data for "FromDate" reduced the total number of records to 32, 640.

Features were split into categorical and numerical features based on their data type and the number of unique features in the feature.

Data Exploration and Visualization

Since the main aim of the analysis is to reduce room cancellations and improve performance, the feature "Booking Status" is used as our target label which has two unique values "Arrived" and "Canceled". 67.2% of the records in the data arrived for their bookings while 38.2% canceled their bookings. This high disparity between the classes in this column shows a class Imbalance which tells us that metrics such as accuracy will be ineffective for evaluating an ML model (Figure 1).

Next, we explored the effect of categorical features on the target label i.e Booking Status. Neptune rooms received the most bookings with a corresponding high amount of cancellations compared to other room types (Figure 2). In Figure 3, we see that most bookings were made for 2 adults for a room which may mean couples or friends.

Figure 1: Distribution of Booking Status

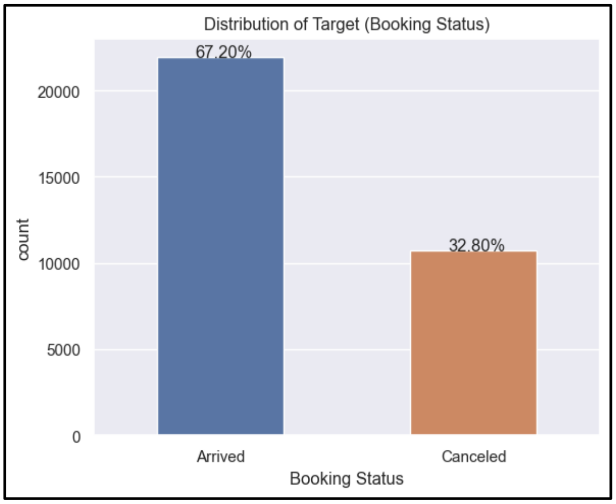


Figure 2: RoomType Vs Booking Status

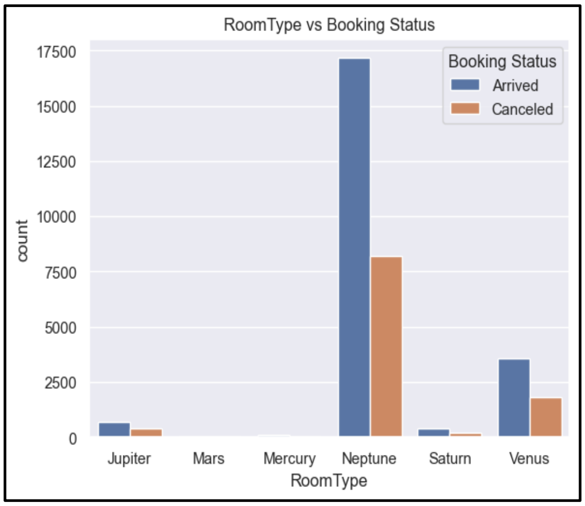


Figure 3: Adults vs Booking Status

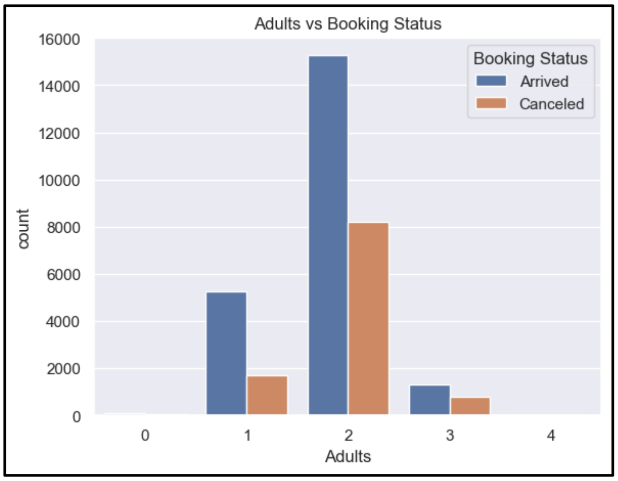
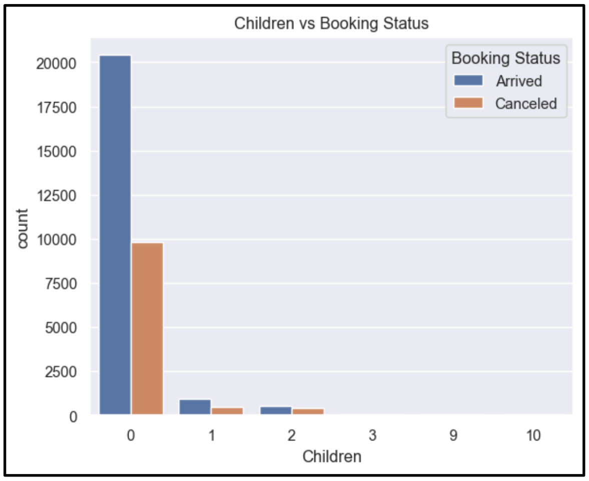


Figure 4: Children vs Booking Status



In Figure 4, we see that most bookings had no children which is not surprising as Figure 3 showed the most bookings were for 2 adults. Most bookings were for Buffet Breakfast than any of the other types of Breakfast. We see zero or little bookings for the Super VIP Breakfast (Figure 5). Figure 6 shows a very interesting insight. As the number of special requests increased, the number of cancellations reduced. Most of the bookings provided 0 special requests which might mean the customers had no idea that they could make special requests. All the bookings which required Airport pickups arrived i.e no cancellations for all airport pickups (Figure 7). In Figure 8, all complementary bookings arrived and most bookings were made Online with very high cancellations. There are few or no aviation bookings. All existing customers arrived for their bookings (Figure 9).

Figure 5: Breakfast vs Booking Status

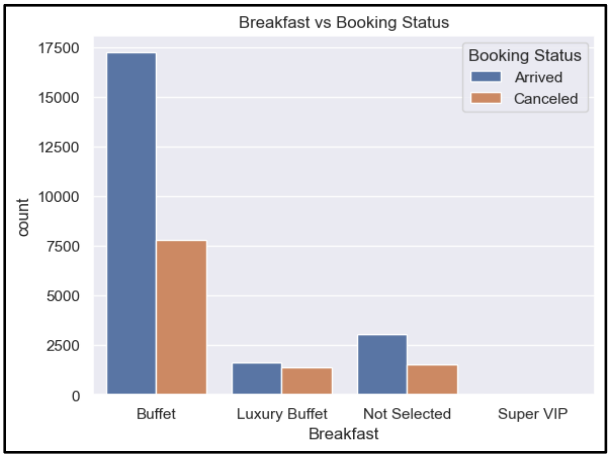


Figure 6: Number of Requests vs Booking Status

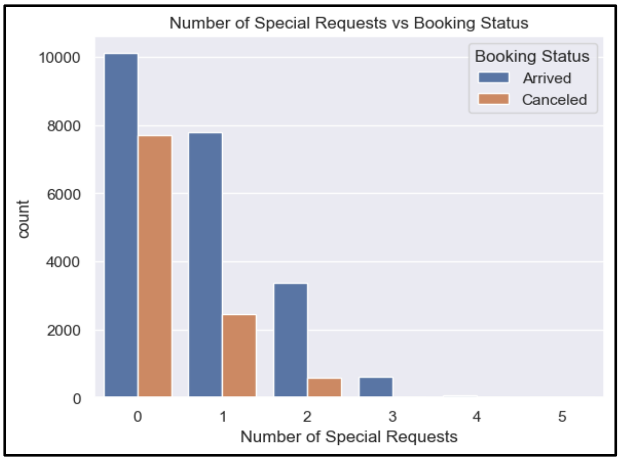


Figure 7: Airport Pickup vs Booking Status

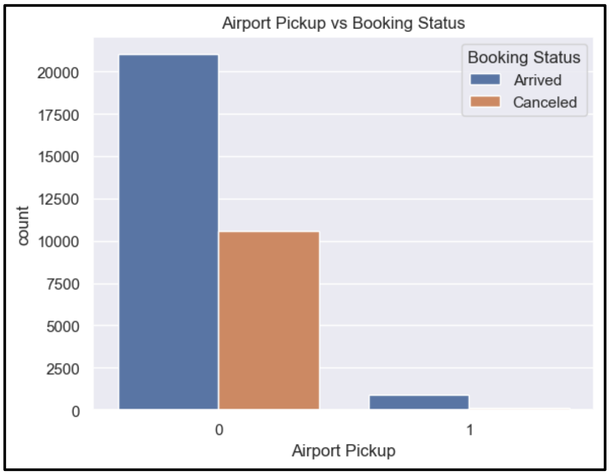


Figure 8: Booking Type vs Booking Status

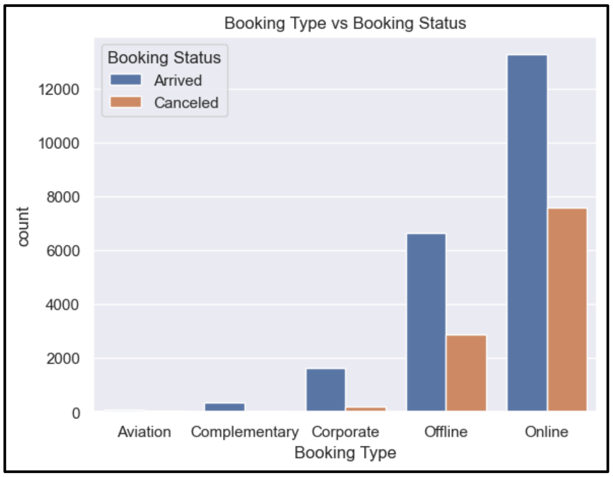
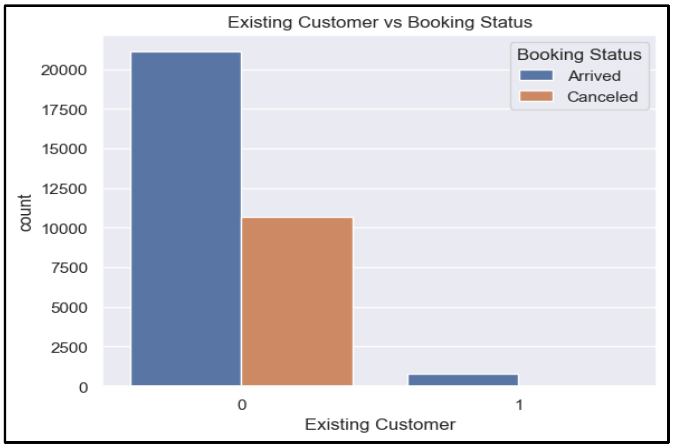
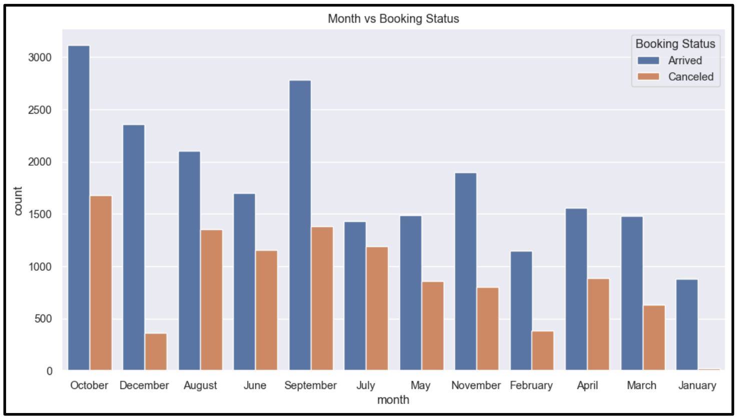


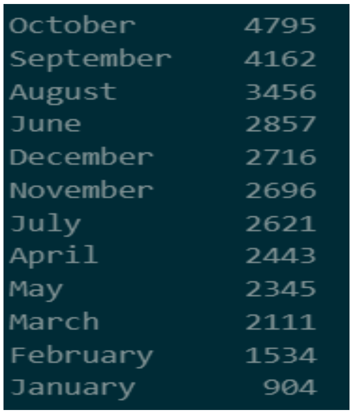
Figure 9: Existing Customer vs Booking Status



Month, day and year were extracted from the “FromDate” column for easier analysis and model prediction. We see that most bookings were made in October, September, August and June. This is not surprising as Labour day is in October, the king’s birthday in June and other public holidays spread across these months (Figure 10). In Figure 11, most bookings were made on Saturdays and Sundays with Thursdays and Fridays having the least number of cancellations.

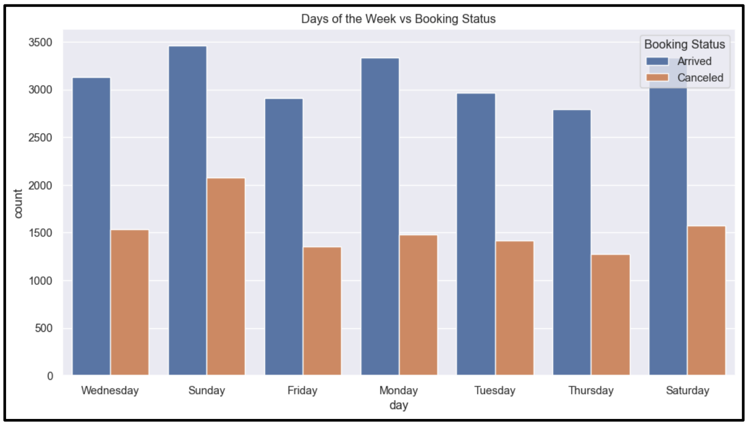
Figure 10: Month vs Booking Status

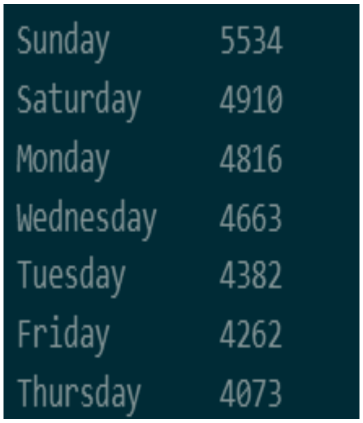




```
October      4795
September    4162
August       3456
June         2857
December     2716
November     2696
July         2621
April        2443
May          2345
March        2111
February     1534
January       904
```

Figure 11: Day of the Week vs Booking Status





```
Sunday       5534
Saturday     4910
Monday       4816
Wednesday    4663
Tuesday      4382
Friday       4262
Thursday     4073
```

Next, we explored the numerical features and how they affect the Booking Status. The distribution of the Average Room Price shows it is almost symmetrical. Using the skew function, it had a skew value of 0.686(Figure 12). Figure 13 shows the distribution of lead time in days and the distribution shows the long tail. The lead time in day had a skewness of 1.289. These skewness are not really high so we used the standard scaler to scale them for the ML model.

Figure 12: Average Room Price

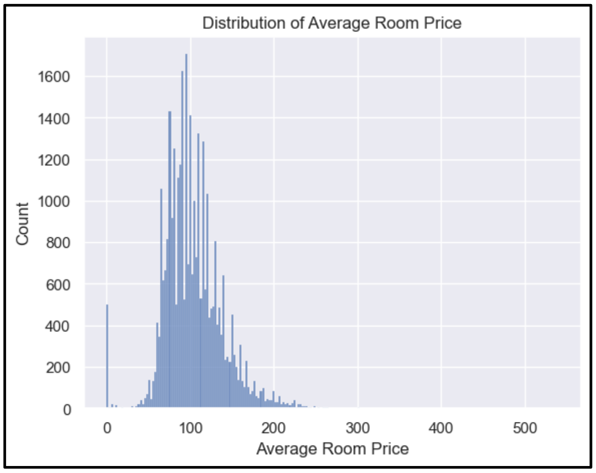
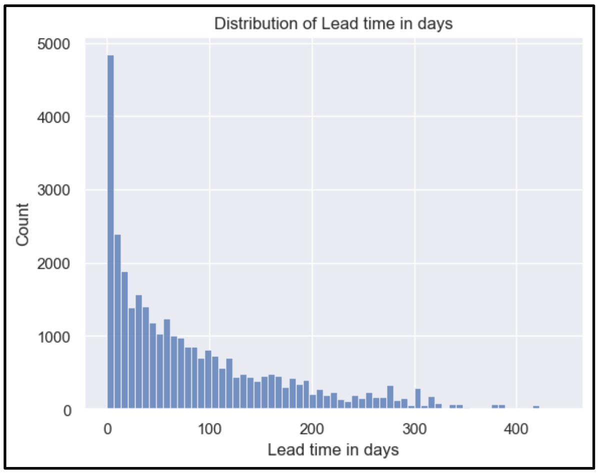
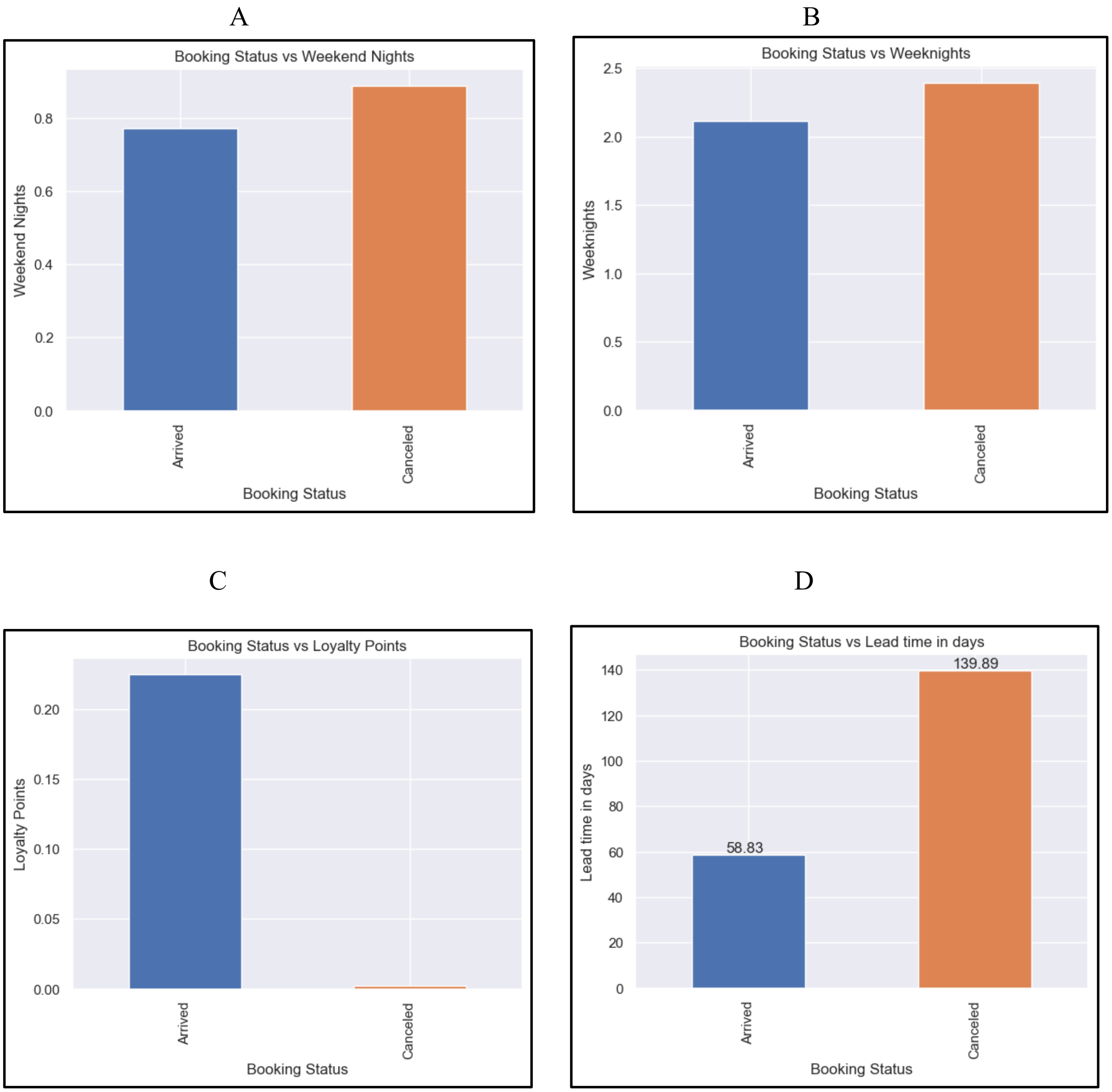


Figure 13: Lead time in days



From Figure 14, we see that features like Weekend Nights(14A), Weeknights(14B), Lead time in days(14D), and Average Room Price(14F) had higher mean values for canceled bookings than arrived bookings. A feature worth looking at is the lead time in days (Figure 14D) which shows the disparity between the mean value of canceled bookings(139.89) and the mean value of arrived bookings(58.89). This shows that arrived bookings had lower lead time in days than canceled bookings. We also see that canceled bookings had little or no loyalty points when compared to that of arrived bookings (14C).

Figure 14: Numericals vs Booking Status



In Figure 15, we see the correlation heatmap of numerical features in the dataset. Notable correlations are between Loyalty points and previous cancellations, Booking Status and Lead time in days.

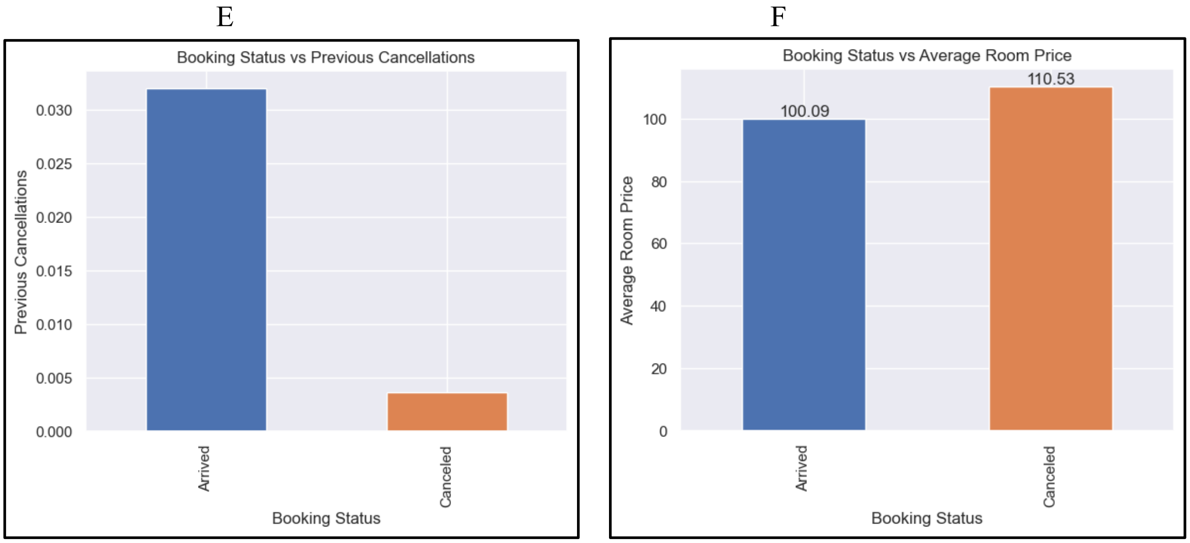
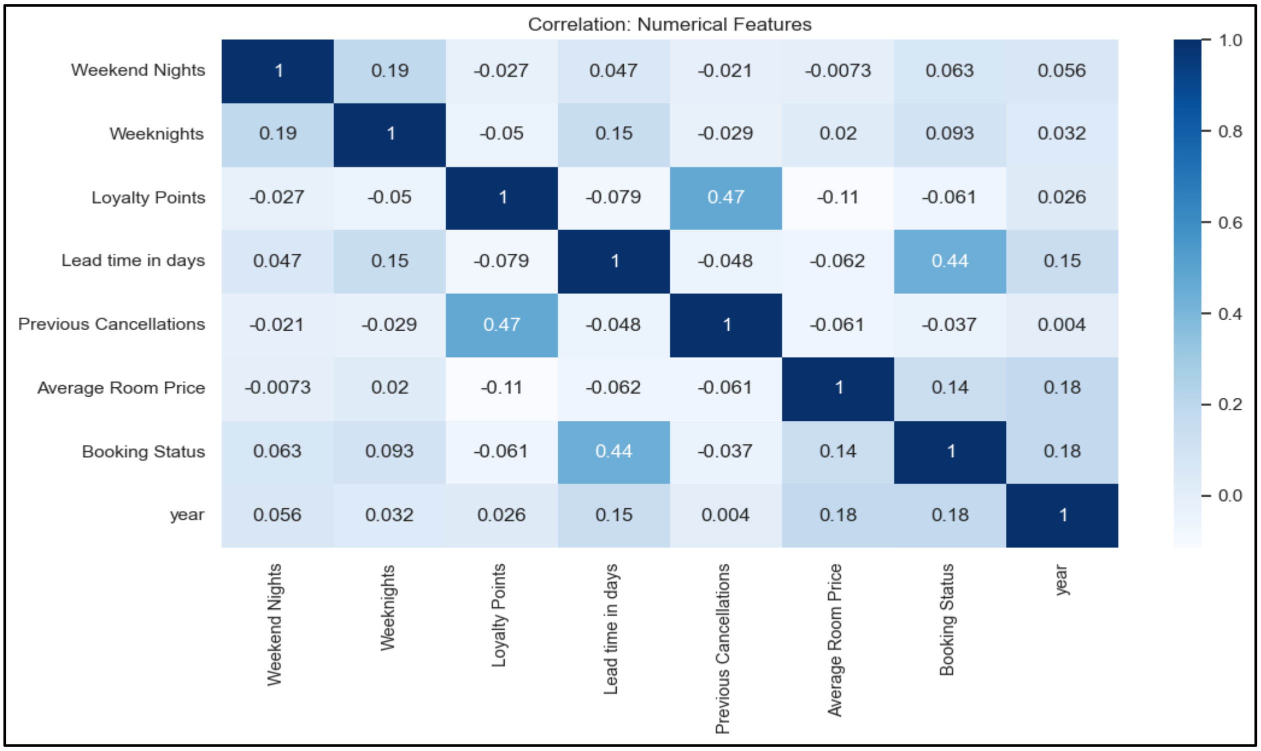


Figure 15: Correlation

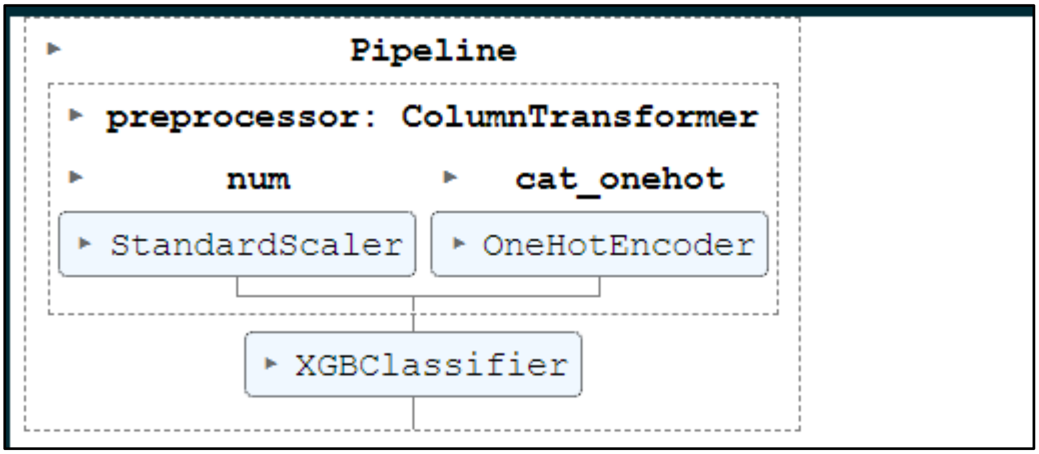


Machine Learning Approach

To make the most of the data provided and develop the best model, several steps were carried out which I call the "Approach". Most of these steps were carried out using a pipeline. The steps include:

- Feature Selection: Features such as BookingID and FromDate were removed from the data used to train the final model to reduce noise the model might experience. The other remaining features were used to train the model.
- Data Preprocessing: First, the target label was encoded by representing "Canceled" as 1 and "Arrived" as 0. Next, features such as month and day were encoded in order. Finally, a preprocessor was created in the ML pipeline that uses a ColumnTransformer to apply different transformers to specific columns. The transformer applies one-hot encoding to the categorical features and standard scaling to the numerical features.
- Model Selection: Three machine learning models, Random Forest, Gradient Boosting, and XGBoost, are trained on the preprocessed data, and their performance is evaluated using the AUC score. The best model is selected based on speed and AUC score. This model is the XGBoost model with an AUC of approximately 0.95 and speed of 2 secs. Next this best model's hyperparameter was tuned using the GridSearchCV and an AUC score of approximately 0.95 was observed.
- Model evaluation: The pipeline uses the AUC score as the evaluation metric for model performance. This metric is appropriate for imbalanced datasets and provides a measure of the model's ability to discriminate between the positive and negative classes. Also, classification reports and confusion matrix for this model were also used to evaluate the model.

Figure 16: Pipeline



Overall, this pipeline provides a comprehensive approach to building and evaluating classification models for imbalanced datasets. By incorporating multiple preprocessing steps, feature selection, resampling, and model selection, the pipeline helps to improve the accuracy and generalizability of the final model.

Model and Performance Metrics

The pipeline used a combination of data preprocessing techniques and machine learning models to classify bookings of Vera Selection Resort (VSR) as potential canceled bookings.

The data was first preprocessed using OneHotEncoder to convert categorical variables to numerical ones. Then, a ColumnTransformer was used to apply different transformations to specific columns in the data. The resulting data was then scaled using StandardScaler to ensure that all features are on the same scale.

Three different machine learning models were used in the pipeline: Random Forest, Gradient Boosting and XGBoost. The performance metric used was area under the ROC curve (AUC) because of the class imbalance in the target label.

Figure 16 : ROC_AUC Curve

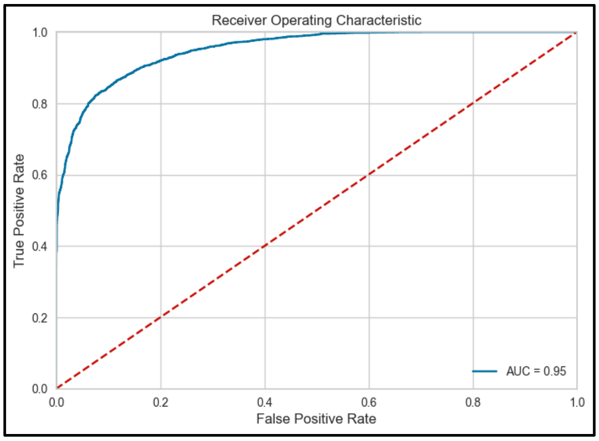
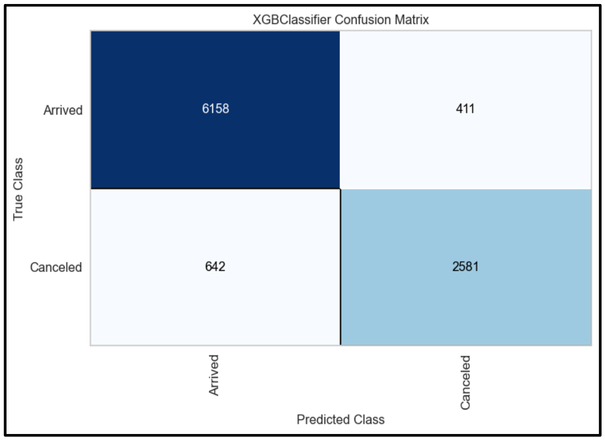
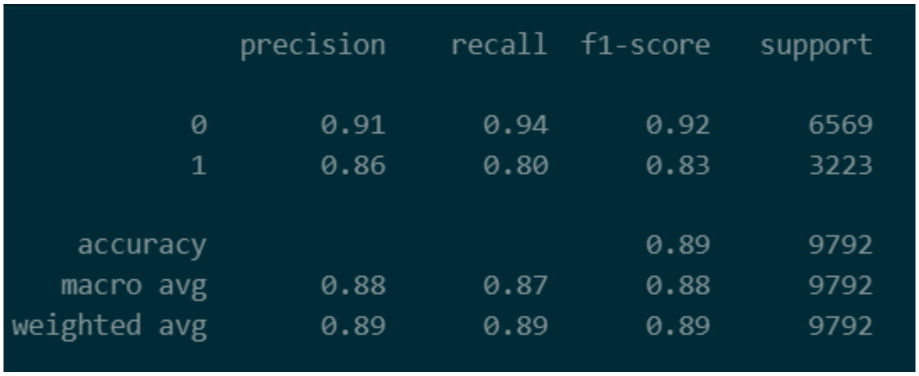


Figure 17: Confusion Matrix



The final XGBoost model had the best performance, achieving an AUC score of 0.95 and an accuracy of 89% on the test set. This indicates that the model can accurately distinguish between canceled bookings and arrived bookings.

Figure 18: Classification Report



```
              precision    recall  f1-score   support

           0       0.91      0.94      0.92      6569
           1       0.86      0.80      0.83      3223

    accuracy                           0.89      9792
   macro avg       0.88      0.87      0.88      9792
weighted avg       0.89      0.89      0.89      9792
```

The confusion matrix showed that the model had a high true positive rate (TPR) of 0.80 and a low false positive rate (FPR) , indicating that the model can correctly identify most canceled bookings while keeping false positives low.

Pros & Cons of the model

XGBoost is a popular gradient boosting library that has become a standard tool in many machine learning applications. Here are some of the pros and cons of using XGBoost:

Pros:

- XGBoost is known for its high accuracy and has been used to win several machine learning competitions.
- It is optimized for both speed and performance, making it efficient for handling large datasets.
- XGBoost has a built-in mechanism for handling missing data and can handle a mixture of continuous and categorical features.
- It can handle a wide variety of objective functions and has a flexible API that makes it easy to customize for specific tasks.
- XGBoost can handle imbalance in the data, making it useful for classification tasks with uneven class distribution.

Cons:

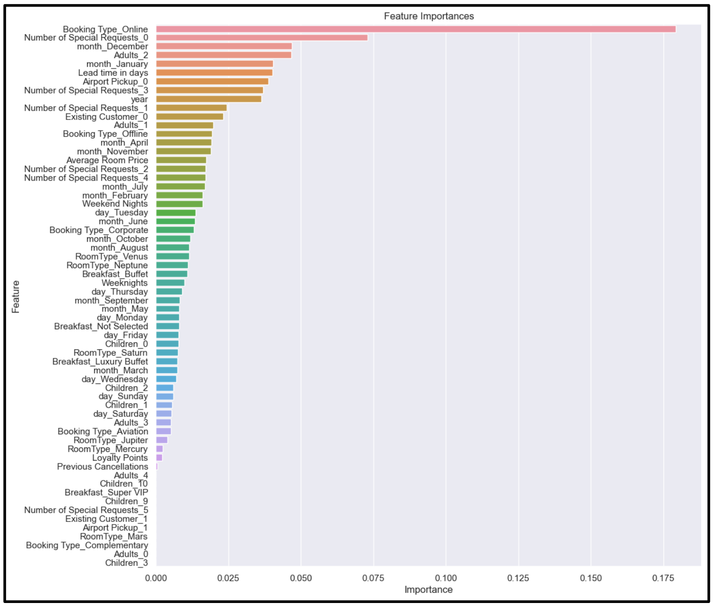
- XGBoost can be computationally expensive and requires tuning of hyperparameters for optimal performance.
- It can be prone to overfitting if not properly tuned and regularized.
- XGBoost requires careful feature engineering to achieve optimal results.
- The interpretability of XGBoost models can be limited compared to simpler models like decision trees or linear regression.
- It may not be the best option for very small datasets, where simpler models may be more appropriate.
- Overall, XGBoost is a powerful and versatile tool for many machine learning tasks, but careful tuning and feature engineering are required to get the best performance.

Business Solution & Recommendations

After analyzing the given dataset and developing machine learning models, our team has identified several key findings and recommendations for Vera Selection Resort (VSR) to reduce room cancellations:

- ➔ Feature Importance: Our analysis shows that the top features contributing to cancellations include online booking, number of special requests, number of adults, month of booking lead time, and airport pickup. We recommend that VSR focus on these features to reduce cancellations.

Figure 19; Feature Importance



- ➔ Machine Learning Models: We tested several machine learning models and found that the XGBoost model outperformed others with an AUC score of 0.95. We recommend that VSR deploy this model to predict cancellations and prioritize customer retention strategies.

- ➔ Early Warning System: We recommend that VSR develop an early warning system that flags bookings with a high probability of cancellation based on the model's predictions. This will allow VSR to proactively reach out to customers and provide incentives to prevent cancellations.

- ➔ Room Booking Policies: We recommend that VSR review and potentially revise their room booking policies to make them more customer friendly. This could include policies around cancellations, refunds, and booking modifications.

Overall, by focusing on the top contributing features, deploying the XGBoost model, implementing an early warning system, engaging with customers personally, and revising room booking policies, VSR can reduce room cancellations and improve customer satisfaction.

References