Hotel Booking Cancellation Prediction

by

Solar Shao, Hongfei Guo

April, 2023

1. Introduction and Motivation

In the highly competitive hospitality industry, managing room inventory is essential to optimizing revenue and profitability. One of the key challenges in managing room inventory is predicting hotel reservation cancellations. When customers cancel their bookings, hotels are left with empty rooms, resulting in lost revenue and operational inefficiencies. This report aims to explore the various factors related to hotel reservation cancellations and how predicting these cancellations can improve yield management in the hotel industry. The literature suggests that several factors are related to hotel reservation cancellations, including lead time, price, seasonality, booking channel, purpose of travel, and length of stay. Here are four reasons why we want to predict hotel reservation cancellations. They are optimizing room inventory, reducing operational costs, improving customer satisfaction, and enhancing pricing strategy. Firstly, it allows hotels to optimize their room inventory by better managing cancellations and ensuring that they are not left with too many empty rooms. It also can increase revenue by selling rooms that would have otherwise gone unoccupied. Secondly, predicting cancellations can reduce operational costs associated with managing overbooked situations, such as staff overtime and delays in room cleaning. Thirdly, predicting cancellations can enhance customer satisfaction by providing more desirable rooms. Lastly, predicting cancellations can allow hotels to develop a more effective pricing strategy by understanding cancellation trends and adjusting rates accordingly

2. Data

2.1 Data Source

Datasets were mainly collected from Kaggle. The entire datasets used for this project contains three separated datasets. The main dataset is the “Hotel Booking Demand.csv” dataset. Besides this, other two datasets that added additional information for hotel booking hence expanded the picture were included. They were data recording countries’ GDP and countries’ happiness level for the time period when bookings were made. Adding these two datasets helped us better understand the geographical differences on booking and cancellation, which might contain useful information for business insights.

Hotel Booking Demand

“Hotel Booking Demand” contains booking information for a city hotel and a resort hotel. Some features are the date the reservation was made, the length of the stay, the number of adults and the number of children for the trip, details about special requests for the booking, etc.

Countries_gdp_hist

“Countries_gdp_hist” contains historical GDP data from 1960 to 2021 by country. It also includes details about the country, such as the name of the region, the name of the subregion, the name of the intermediate region and country code, and GDP variation from previous year. Only GDP data from 2015 to 2017 was selected for future uses in this project.

World Happiness data(2015-2017)

“World happiness” contains information about global happiness. Happiness data from 2015 to 2017 was used since hotel booking data only covered data from this period. It includes features such as happiness rank, happiness score, etc. Three years’ happiness data was combined into one dataset using matched columns.

Since all datasets contain country code, which was in ISO 3 digit format, and year columns, they were merged into one using these matched columns.

2.2 Data Visualization

Categorical Data Distribution

Data visualization was done in order to understand the data better. For each categorical variable, distribution plots were drawn to see the distribution among categories(Fig. 1.) From the distribution plot, there was a slight difference between cancellation group and non-cancellation group given that there were 40,457 bookings being canceled and 61,577 bookings not being canceled. It did not seem to be a sign of group imbalance and would not cause issues in modeling. In the dataset, there was about two times more city hotel data than resort hotels. In the next step, we separated the two groups and did visualization on individual groups to exclude this factor. Speaking of year, the number of data from 2015 to 2017 contained in our dataset were 20,017, 33,268 and 48,749, which showed a slight sign of group imbalance and should be awarded in result interpretation.

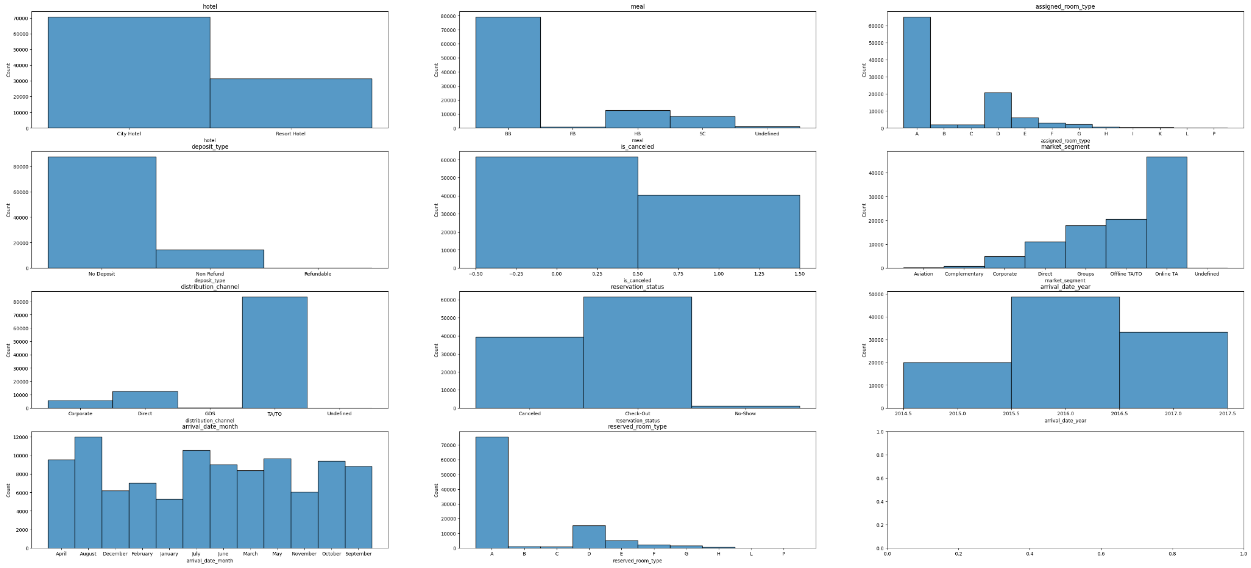


Fig 1. Categorical Data Distribution

Geographical Distribution

To understand where the customer came from, a geographical heat map was drawn(Fig. 2.). From the plot, we could see that customers mostly came from Asia(excluding Russia), South America, Africa, Europe, and Australia. For North America, there were customers coming from Mexico and there was no data for customers coming from Canada and the United States. It may be caused by data exclusion when collecting the data, or simply because there was no customer coming from these countries to live in the two hotels. Since there was no information about data exclusion, we assumed that there was just no customer coming to these two hotels.

From the distribution plot, most of the customers were coming from Europe. From the numerical analysis, most of the customers were coming from Portugal(21,071) and the United Kingdom(8481).

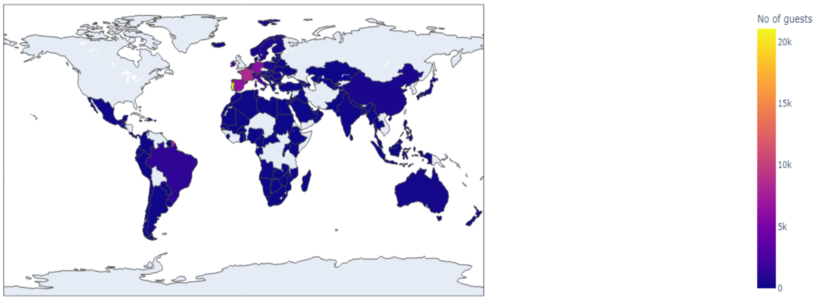


Fig 2. Geographical Distribution

Customer Type Percentage by hotels

Pie charts showing the percentage of customer type for each hotel were drawn(Fig. 3.). Most of the customers were transient guests for both hotels- making reservations on their own and not associated with other transient bookings. Transient-Party customers were the second largest population in the data which were those making bookings associated with other bookings.

Hotel Price Variation by Time

Box plot containing hotel prices for each month and year separated by hotels was drawn(Fig.4. and Fig.5.). Hotel prices were higher and had larger variations in June, July and August when students were in school break.There was also an interesting result that only for these three months, average daily rate for resort hotels was higher than city hotel. Average daily rate for the three years was roughly the same.

Cancellation Variation by Month

To understand how cancellation cases varied by month, histogram about cancellation rate variation separated by month was drawn(Fig.6.). It could be seen that cancellation rates were high from April to October, which were above 38%, and were low in the winter months from November to March. This might be caused by there just being less customers in the winter months or other reasons related to less cancellation actions in the winter.

Number of Guests Variation by Month

From the plot(Fig.7.), there were always more guests in city hotel than resort hotel around the year. For city hotel, the number of guests increased from winter seasons(January) to its peak in the middle of summer(August). Then it started to decrease for the rest of the year. For resort hotel, it had a similar trend but had a decreasing pattern from March to June.

Cancellation Action by Guests Type

To understand the cancellation action between different guests, cancellation rates between repeated guests- guests made reservations to the hotel previously- and non-repeated guests were calculated. The cancellation rate for repeated guests was 14.95% and for non-repeated guests was 40.57%, which was consistent with common sense that it is always harder to gain trust from new customers.

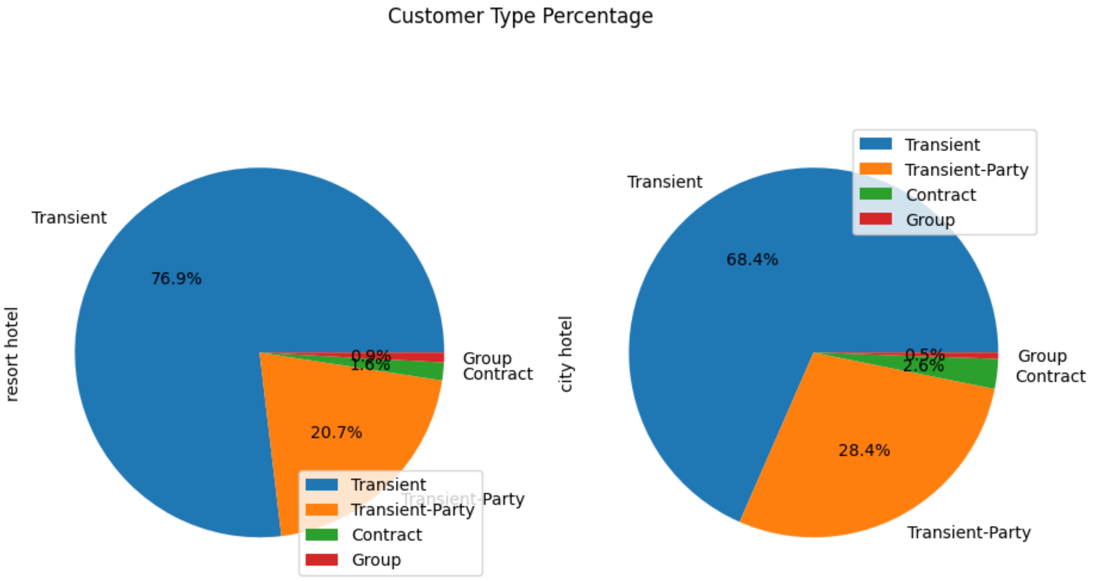


Fig 3. Customer Type Percentage

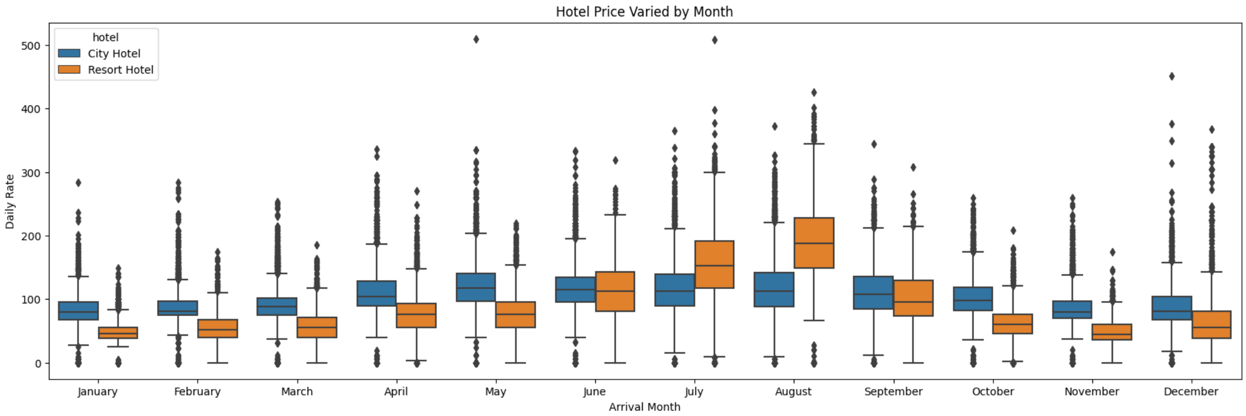


Fig 4. Hotel Price Variation by Month

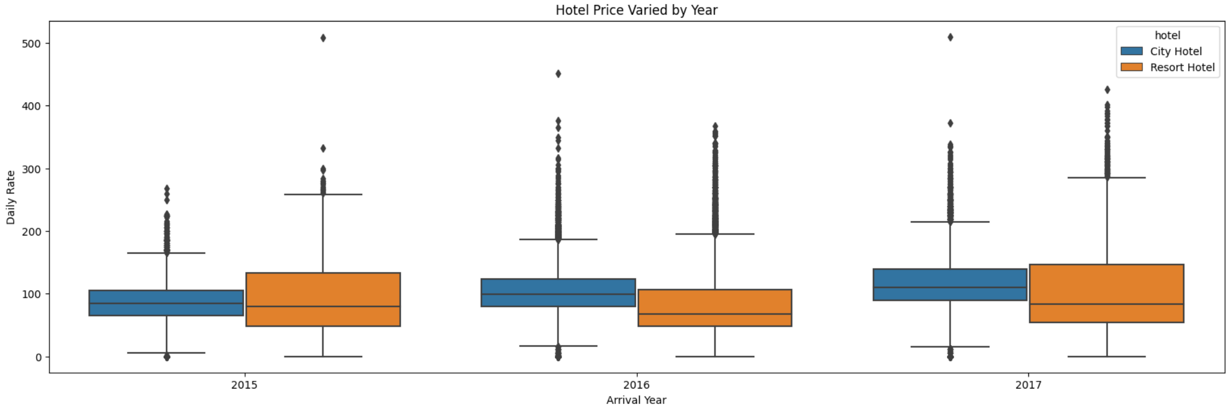


Fig 5. Hotel Price Variation by Year

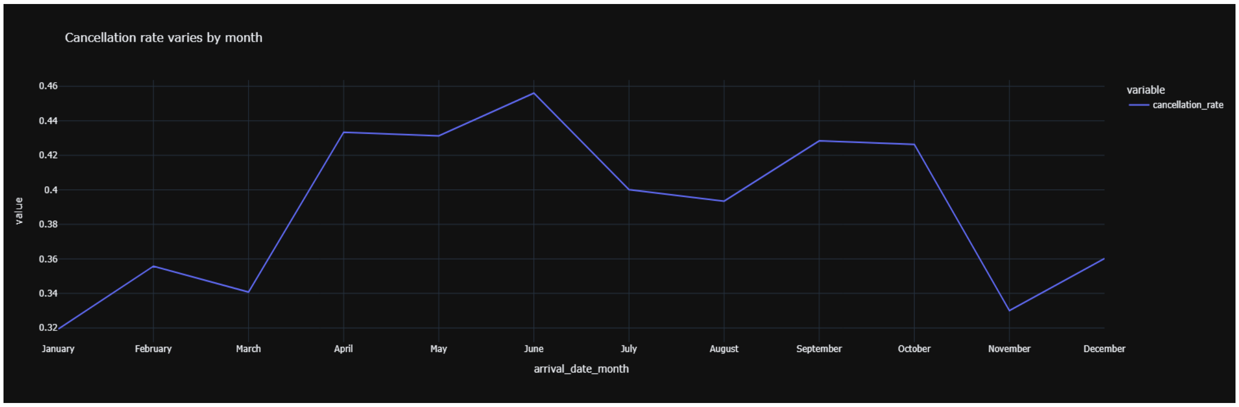


Fig 6. Cancellation Rate Variation by Month

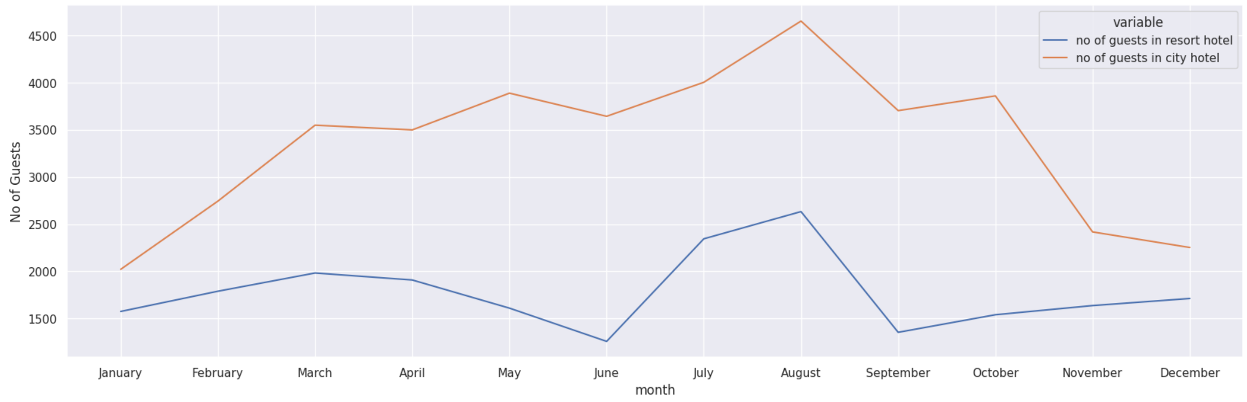


Fig7. Number of Guests Variation by Month

2.3 Data Preprocessing

Data was then preprocessed to make the modeling process easier and the result more accurate. A new column called “same_room_type” was generated using “reserved_room_type” and “assigned_room_type” by comparing whether the customers were assigned to the room they reserved, which was considered important to the cancellation action.

Unnecessary columns were picked and dropped from the original dataset to reduce model complexity and increase accuracy. There were a few examples of unnecessary columns: happiness rank(which was related to happiness score and happiness score was kept in the dataset), arrival_date_year(since there was no global pandemic during the study period, the cancellation action pattern for the three years could be considered the same), total_gdp_million(same as total_gdp and total gdp was kept), reservation_status, regional name, country name, etc. This process reduced the total feature numbers from 50 to 26.

Data was then divided by categorical columns and numerical columns to perform separated transformations. Categorical data was encoded using one-hot encoding. For numerical data, null values were filled using 0. Intended to increase the accuracy of the models and reduce training time, standardization was performed on numerical data. Categorical data and numerical data after preprocessing were then combined back together for modeling.

3. Modeling

In order to predict hotel reservation cancellations, hotels can use a variety of predictive modeling such as data Logistic Regression, Random Forest, XGBoost and Neural Network. By analyzing historical data on cancellation rates and the various factors related to cancellations, hotels can develop algorithms that can predict cancellations with a high degree of accuracy.

3.1 Logistic Regression

Since this is a classification problem, logistic regression is a good model to start with. The function maps input data to the probability of cancellation. And by choosing a threshold(set as 0.5 for this project) as the threshold to conclude whether or not cancellation happens, the model outputs a binary classification result. Pre-built Logistic Regression model in sklearn was implemented for this project.

3.2 Random Forest

For the data set, random forest is a great model to predict cancellation, because random forest can handle both numerical and categorical data effectively. Hotel reservation data includes various types of data such as guest information, room type, booking dates, etc. In addition, random forest handles missing values and captures nonlinear relationships between the features and the target variable. Overfitting occurs when a model is too complex and is fitted too closely to the training data, resulting in poor performance on unseen data. Random forest can reduce overfitting by using multiple decision trees and aggregating their results. During applying the random forest model, By fine-tuning these hyperparameters, we optimized the performance of the random forest model for the specific problem and data set. We set min_samples_leaf = 5 which means that each leaf node in the decision trees of the random forest should have at least 5 samples; n_estimators = 100 which means that the random forest includes 100 decision trees. A larger value for this hyperparameter can improve the accuracy and stability of the model, but it can also increase the computational cost and the risk of overfitting. We applied the most common type of cross-validation, k-fold cross-validation to evaluate the random forest model. We would average the results from the 10 iterations to obtain a final performance estimate.The advantage of using k-fold cross-validation is that it provides a more reliable estimate of the model's performance compared to a simple train-test split. The results are averaged to obtain a more robust estimate of the model's performance.

3.3 XGBoost

By using techniques such as weighted loss functions and subsampling, the XGBoost can effectively address this issue and produce accurate predictions. XGBoost has many hyperparameters that can be tuned to improve model performance. We tuned n_estimators = 100 which means that 100 decision trees were built in the ensemble; learning rate = 0.3. This means that in each iteration of the boosting process, the weights of the residuals are multiplied by 0.3 before they are added to the new tree. We also used 10-fold cross-validation to evaluate the model. Finally, we would average the results from the ten iterations to obtain a final performance estimate.

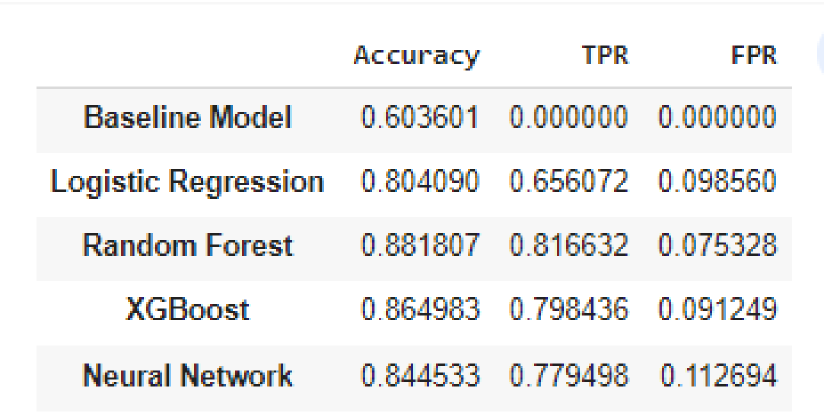
3.4 Neural Network

Since there were 59 features for the input data, a shallow neural network might be able to catch the complicated relationships and perform well. A seven-layer neural network was built to test the hypothesis. The units for the layers were 100, 84, 64, 64, 32, 16 and 1. For the first three layers, “relu” was used as the activation function to add non-linearity to the model hence adding complexity. “Sigmoid” activation was used for the last layer to perform classification. Since this was a classification problem, “binary cross entropy” was used as the loss function to calculate the loss and adjust weights in the training process. “Adam” optimizer was used as it usually reaches the optimal result fast. Accuracy was evaluated while training. Keras was used to build and train the neural network. During training, 33% of the training data was used for validation and early stopping was used to monitor validation loss to avoid overfitting.

4. Result

The model results were shown in Figure 8. Random Forest had the highest testing accuracy, which was 88.18%, followed by XGBoost, which had a testing accuracy of 86.49%. The neural network did not perform better and had an accuracy of 84.45%. The logistic regression had the lowest test accuracy, which was 80.40%. Speaking of other evaluation metrics, Random Forest stood out again with the highest True Positive Rate(TPR) - 0.81 and the lowest False Positive Rate(FPR)- 0.07, followed by XGBoost. Though the FPR for neural networks was higher than logistic regression, neural networks had a much higher accuracy and TPR. Thus, we still considered that neural networks performed better than logistic regression. From the accuracy and loss trend plot for neural networks, the training procedure stopped as the validation loss went down as we set up early stopping when fitting the model.

Random Forest was selected as the final model for this problem due to its high accuracy and interpretability. Using written in sklearn random forest, feature importance calculated using mean decrease in impurity was shown in Figure 11. For the decision tree, the best features that minimize the impurity of the resulting subsets were selected to partition the tree at each middle-step. Since Random Forests was a model combining multiple decision trees together with randomness added, the features with larger decrease in impurity would be more important. The top five most important features related to cancellation action were: The Hotel Daily Rate, Deposit type(non refund), Happiness Score, Deposit type(no deposit), Arrival date day of month.



	Accuracy	TPR	FPR
Baseline Model	0.603601	0.000000	0.000000
Logistic Regression	0.804090	0.656072	0.098560
Random Forest	0.881807	0.816632	0.075328
XGBoost	0.864983	0.798436	0.091249
Neural Network	0.844533	0.779498	0.112694

Fig 8 Evaluation Table for Models

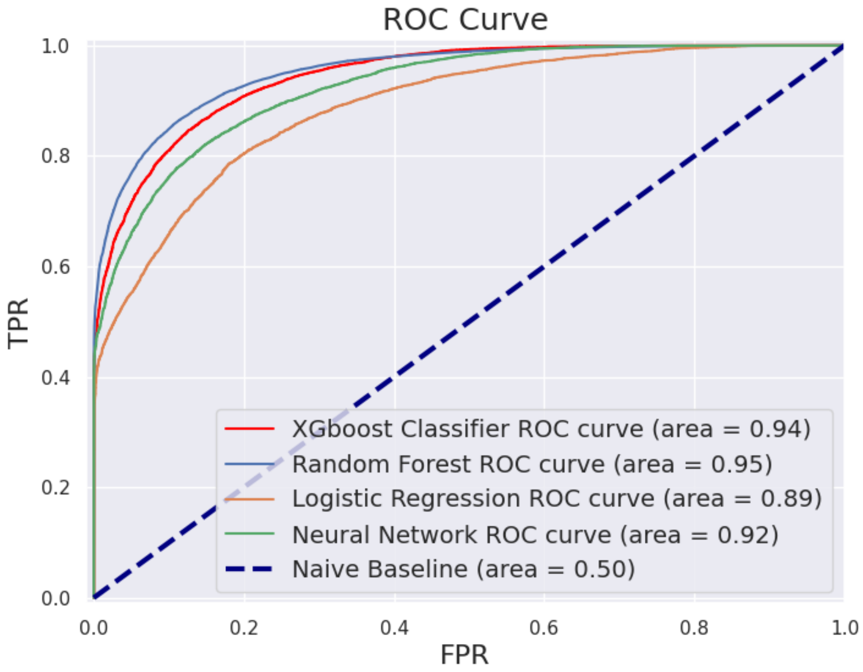


Fig 9. ROC curve for models

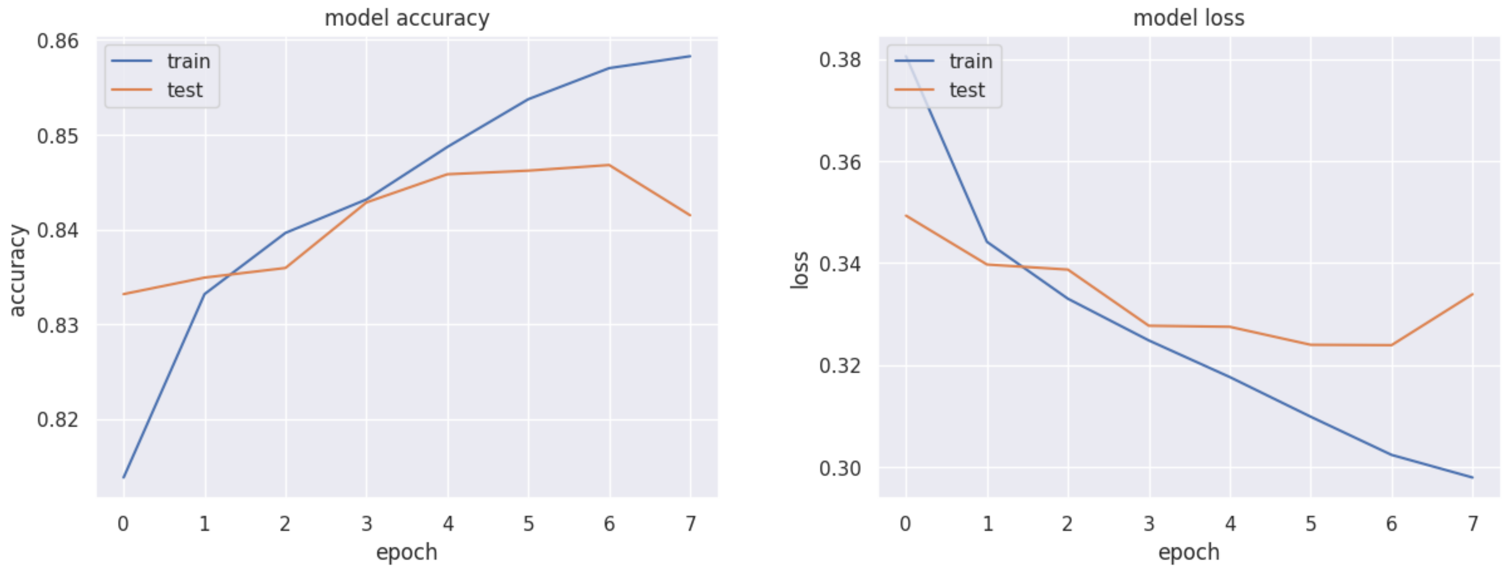


Fig 10. Neural Network Model Accuracy and Loss Trend by Epoch

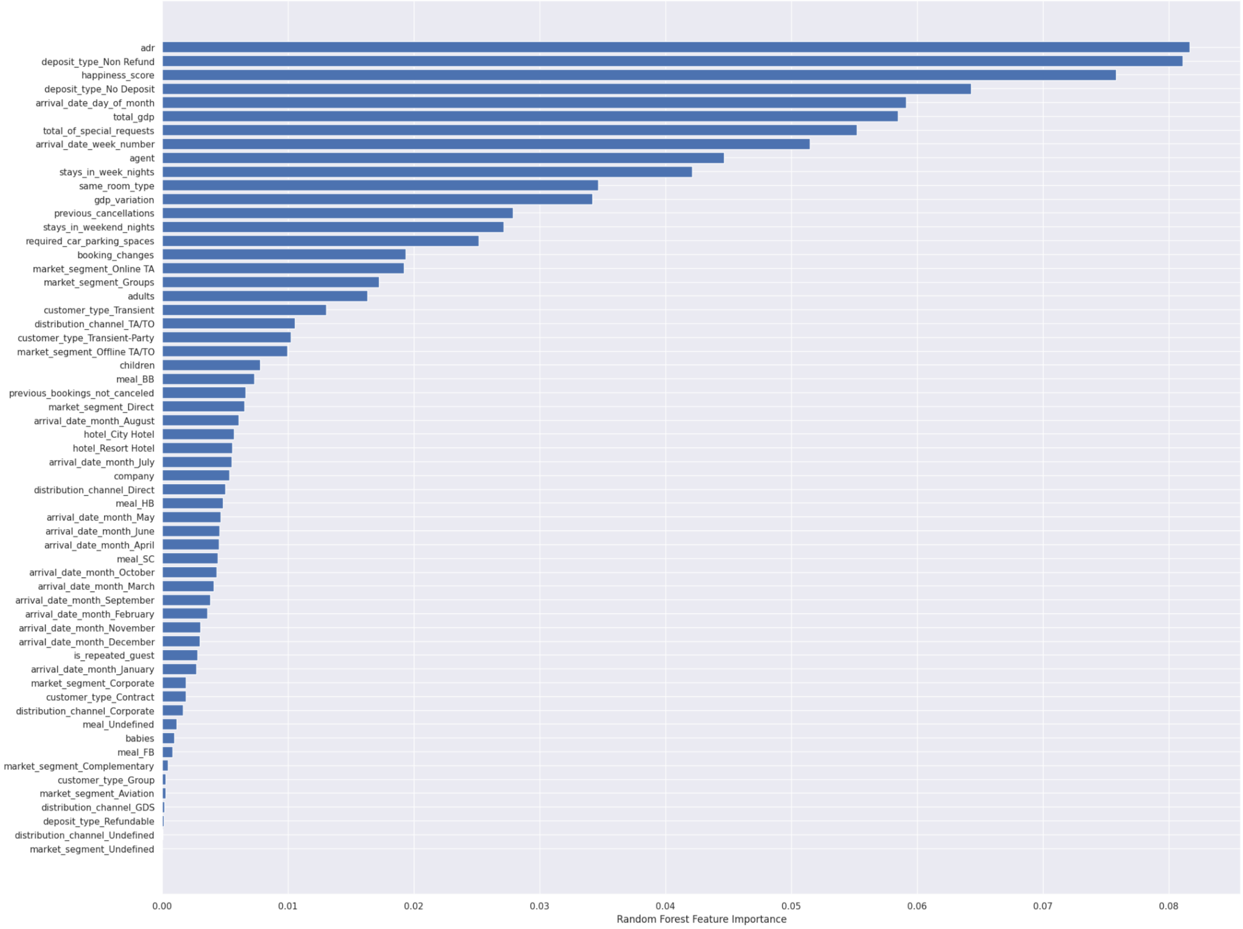


Fig 11. Feature Importance for Random Forest Model

5. Impact and Conclusion

To further improve the current result, we could implement cross-validation to select the best parameters for a random forest model, such as n_estimators, nax_leaf_nodes. We could use bootstrap to construct confidence intervals of OSR2. The bootstrap is a powerful technique for estimating the variability of performance metrics. The prevalence of the bootstrap is due in large part to increases in computational capabilities. The hotel yield management may be significantly impacted by using a machine learning model to forecast hotel reservation cancellations. Hotels can manage their inventory and modify their pricing tactics to avoid revenue loss by anticipating cancellations. Additionally, anticipating cancellations can help hotels manage their staffing needs and resources more effectively. To broaden the analysis's scope and enhance its impact, some related variables could be taken into account, such as weather forecasts, local events, client demographics, or US and Canada booking information. For instance, if a hotel is situated in a region with a high probability of bad weather, adding meteorological data to the model can improve the prediction of cancellations. Incorporating information about regional events, like concerts or conferences, can help predict cancellations due to changes in travel plans. The impact of the model may vary across different subpopulations of interest. For instance, the model might have a greater impact on budget hotels with higher cancellation rates than on luxury hotels where cancellations are less frequent. Additionally, the impact of the model may differ across different types of customers, such as business travelers versus leisure travelers. While using machine learning models to forecast hotel reservation cancellations has many potential advantages, there are some drawbacks to take into account as well. If hotels use these predictions to overbook their rooms, it could result in some customers being turned away upon arrival. This could damage the hotel's reputation and lead to negative reviews or even legal action. Additionally, customers may be hesitant to make reservations at a hotel that is known for overbooking, leading to lost revenue in the long term. Overall, appling a machine learning model to forecast hotel reservation cancellations has potential benefits, but it's crucial to carefully consider the potential negative consequences and to balance the benefits with the risks.

6. Reference

Nuno Antonio, Ana de Almeida, Luis Nunes, Hotel booking demand datasets, Data in Brief, vol. 22, 2019, Pp. 41-49, ISSN 2352-3409,
https://doi.org/10.1016/j.dib.2018.11.126.
(https://www.sciencedirect.com/science/article/pii/S2352340918315191)