Business Understanding

Our client is a city hotel located in Lisbon, Portugal. We have identified a significant business problem: the hotel has a high cancellation rate of 41%. This issue has a multifaceted impact on the hotel's overall profitability and revenue management. High cancellation directly leads to revenue loss especially when the cancellations occur close to the check-in date. It also causes difficulties and inefficiencies in the hotel's operation and financial budget planning (Duetto, 2016).

To address this issue, we will develop a predictive model to help the hotel anticipate the likelihood of a reservation being canceled. By using this model, the hotel is able to deploy optimal revenue management techniques such as overbooking and pricing strategy to mitigate the issue of jeopardized profitability and operational inefficiency. Specifically, this model enables the hotel to use the overbooking technique more effectively. It can help to avoid excessive overbooking, which leads to guest dissatisfaction and compensation costs to some extent. Hence, the hotel could improve profitability without compromising extra costs incurred and reputation damages.

Uncertainty: If the reservation is going to be canceled

Decision / deployment: Implement overbooking for the reservation (O) or Not implement (NO)

Goal: max E[Profit | X, O]

Expected value of decisions:

E[Profit | X, O] = P (canceled | X, O) * E[V(X,O) | canceled, X, O]

E[Profit | X, O] = P (NOTcanceled | X, O) * E[V(X,O) - c | NOTcanceled, X, O]

* c is the compensation costs for overbooking guests

Data Understanding

We obtained our dataset "Hotel Booking Demand" from Kaggle.com, which originated from the article *Hotel Booking Demand Datasets*, written by Nuno Antonio, et al. for Data in Brief, Volume 22, February 2019. To be specific, it contains booking information from a resort hotel (*H1*) and a city hotel (*H2*), and some corresponding details such as the reservation time, length of stay, the reserved room type, whether the reservation is canceled, the customers' demographics, etc. Both datasets share the same structure, with 31 variables describing the 40,060 observations of *H1* and 79,330 observations of *H2*. Also, all information regarding customers' identification is deleted for data and privacy protection. This dataset involves some potential biases. One of which might be temporal bias because the information was collected before COVID, so the models and evaluations we generated for now might not be appropriately applied in the post-COVID period as people's hotel reservation and cancellation pattern might have changed with carry-over effects from the pandemic.

Data Preparation

Our data preparation involves two major types, data removal and data standardization. In general, after cleaning the data, we finalized the dataset with 14 variables and 79,325 observations for later data modeling. To be specific, *is_canceled* is our target variable as well as the dependent variable, while all other variables should be the independent variable.The data dictionary is attached as Appendix A.

1. **Data Removal:**

1) Irrelevant Variables: Since we are specifically targeting the city hotel, we dropped the unwanted rows of resort hotels, which were stored as *H1* under the *hotel* column. Then, we removed the entire *hotel* column.

2) Ambiguous Variables: *reserved_room_type* and *assigned_room_type* were dropped because there is no information about the differences between room types which makes it unable to clarify the relationship between these two variables and the target variable.

3) NULL values: *agent* and *company* columns were dropped since there are too many NULL values.

4) Hard-to-categorize Variables: We also dropped the *country* column because there are 167 distinct countries which might later generate too many dummy variables, and makes it too complicated for modeling.

5) Duplicate Variables: Since our target variable *is_canceled* is already the dummy variable representing the meaning of *reservation_status*, we only need to keep *is_canceled.*

2. Data Standardization:

1) Conversion to Numerical Variable: *arrival_date_month* (e.g. July to 7).

2) Conversion to Dummy Variable: *market_segment, distribution channel, deposit_type, customer_type.*

3) Adding New Dummy Variable: since we dropped the two variables related to room types, which may lead to some information loss, we added a new dummy variable to compare whether *reserved_room_type* and *assigned_room_type* are the same (1 stands for difference, 0 stands for consistency) and we have checked the correlation between this variable and our target variable, which is 0.213 and larger than our benchmark 0.05.

Modeling

1. KNN

We first ran the KNN model because it is intuitive and simple. It also has the advantage of capturing localized patterns in the data. Hotel booking patterns can exhibit temporal variability. Capturing local patterns helps the model adapt to short-term fluctuations such as seasonal trends, holidays, or special events that may affect cancellation behavior.

However, KNN is limited since it relies on calculating the distance between data points, the computational cost becomes high for larger datasets. In addition, its performance may degrade when dealing with high-dimensional data. Whereas, our data has 79325 rows and 59 features after one thermal encoding.

Therefore, we decided to use **Lasso** as a feature selection method to reduce the number of features. After feature selection, we reduced the number of features from 59 to 13. Then we applied KNN to see the result and the result is much better than before. However, 13 features are still too many for KNN and the result is not good enough. Therefore, we also tried a logistic regression model.

2. Logistic Regression

Secondly, we choose logistic regression because of its computational efficiency and interpretability. It is computationally efficient, fast to train. This efficiency is very favorable when working with large datasets or when real-time or near real-time predictions are required. Logistic regression is also well suited for risk assessment tasks such as predicting the probability of a binary outcome. In the case of hotel room cancellations, knowing the probability of cancellation is critical for proactive revenue management and strategic decision making.

Logistic regression also has its limitations. It has a limited ability to model nonlinear relationships, which can be a disadvantage when dealing with complex interactions in the data. Moreover, it assumes that the independent variables are independent of each other. If there is a strong correlation (multicollinearity) between the predictor variables, it can lead to unstable coefficient estimates that make it difficult to explain the significance of individual features. For example, in our data, *'repeated guest'* and *'previous non canceled'* have a relatively high correlation, which is 0.45. So, we tried the random forest model.

3. Random Forest

Unlike linear models, Random forest models are good at capturing nonlinear relationships and dealing with high-dimensional data, making them ideal for predicting complex hotel room cancellations. Furthermore, random forests are inherently robust to overfitting, especially when compared to individual decision trees. The ensemble nature of random forests combines multiple decision trees to help mitigate overfitting. However, random forests sacrifice some interpretability compared to simpler models. The ensemble nature of random forests makes them less interpretable than simpler models such as logistic regression. It can be challenging to understand the detailed decision-making process for each tree in the forest. Random forests are computationally intensive, especially with large numbers of decision trees and complex data sets. Training a large number of decision trees may require significant computational resources. Although random forest has a high prediction accuracy, its predictions are often considered a "black box". It's not easy to explain the model's decisions to stakeholders or regulators.

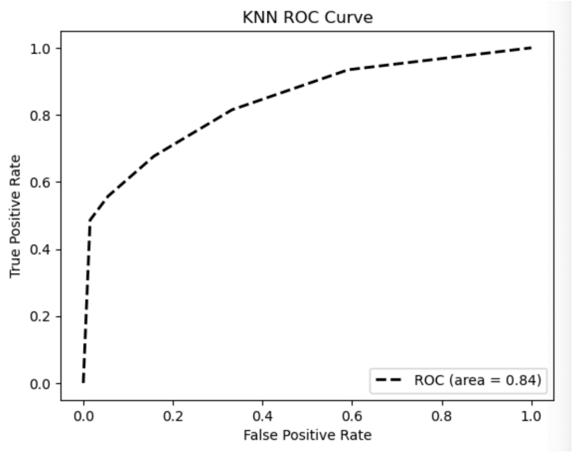
4. Decision Trees

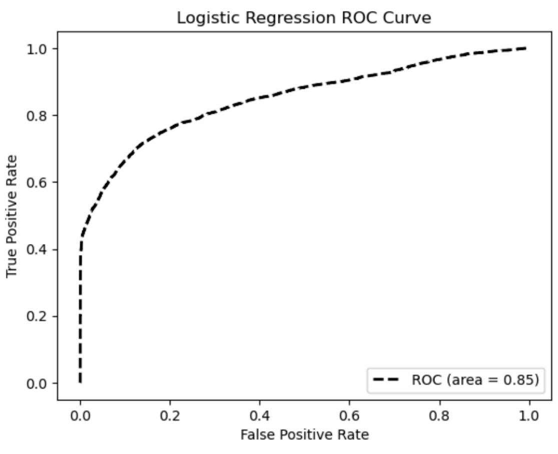
To demonstrate the decision-making process more clearly, we modeled the data using a decision tree algorithm. Decision trees require minimal data preprocessing to clearly demonstrate the decision-making process. Decision trees do not assume a linear relationship between features and target variables. This flexibility helps to deal with complex nonlinear patterns that may affect hotel room cancellations. However, decision trees are prone to overfitting, especially if the model is too complex. In addition, decision trees can be limited in their ability to model complex nonlinear relationships.

Evaluation

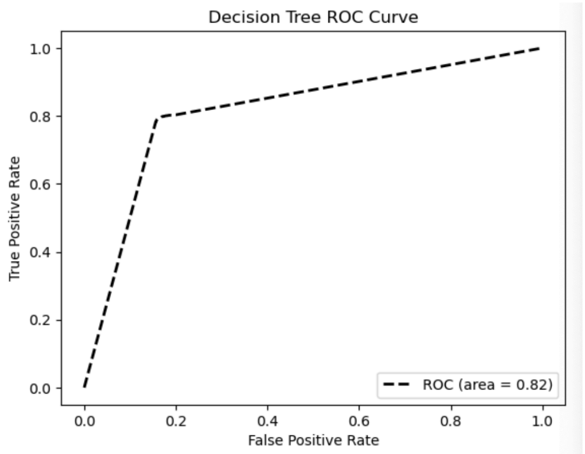
1. Matrix

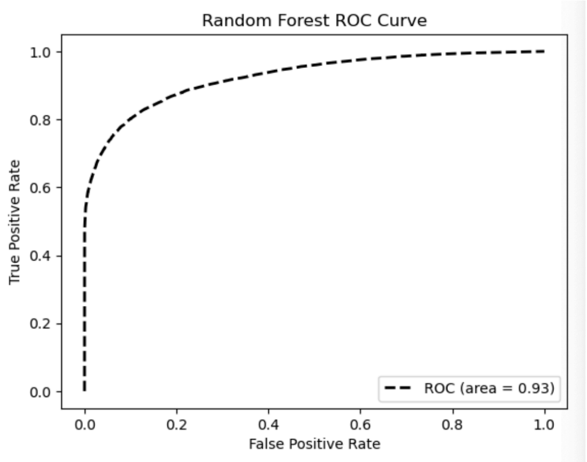
After Lasso feature selection, we first use ROC and AUC as criteria to choose a winning model. The reason is that the ROC and AUC score integrates the performance of the binary classification model under different thresholds, and is an effective metric for comprehensive evaluation independent of category imbalance and threshold selection. Also we use accuracy as a supporting evaluation indicator. Below are the visualizations of our models' performance.

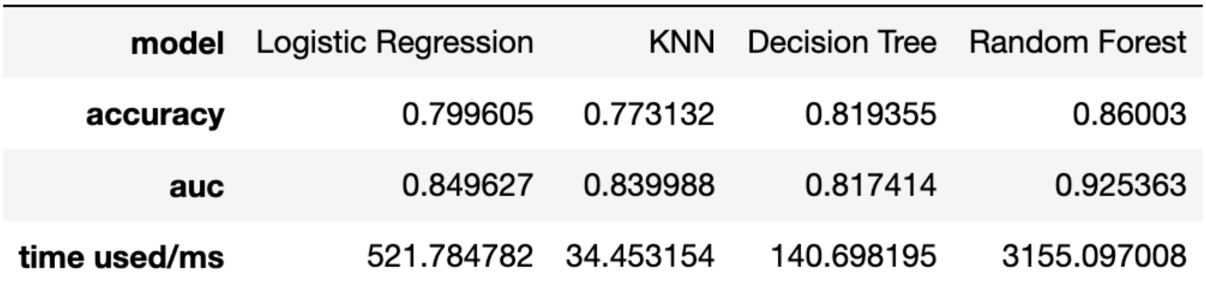




.







model	Logistic Regression	KNN	Decision Tree	Random Forest
accuracy	0.799605	0.773132	0.819355	0.86003
auc	0.849627	0.839988	0.817414	0.925363
time used/ms	521.784782	34.453154	140.698195	3155.097008

Combined with ROC analysis, including AUC, it seems that the Random Forest Model is the best model among these four.

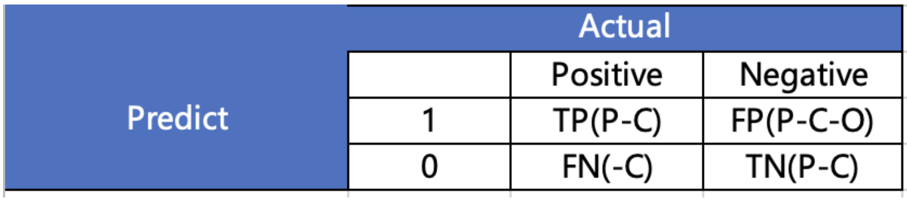
2. Reselection with Expected Gain

Above the goal of cancellation estimation, improving the client's benefit should be considered with real business context. Reselection will be done with calculating expected gain, based on confusion matrix, which will be influenced by factors below:

C: Operating cost per room: the average expected cost for each room (C>0).

P: Price per room: the expected price from one successful check-in (P>0).

O: Offset per overbooking: the expected compensation that the hotel needs to pay for its customer if they cannot offer a room because of overbooking (O>0), for example, they need to help consumers find a nearby hotel and pay for the room.



		Actual	
Predict		Positive	Negative
	1	TP(P-C)	FP(P-C-O)
	0	FN(-C)	TN(P-C)

Income without prediction:

$$Income1 = (TP+FN)\times(-C)+(FP+TN)\times(P-C)$$

Apply the overbooking strategy if the room is predicted to cancel:

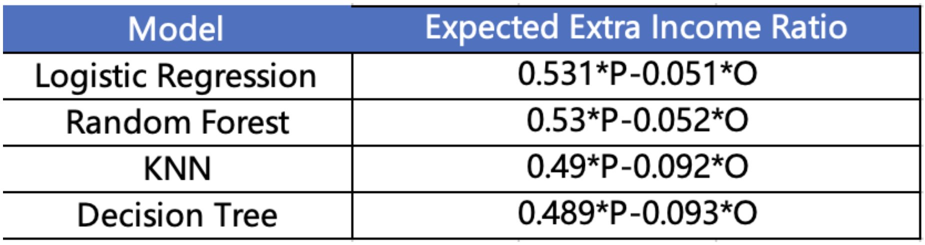
$$Income2 = TP\times(P-C)+FN\times(-C)+FP\times(P-C-O)+TN(P-C)$$

Extra income with prediction:

$$Income2-1 = B-A/\mathrm{n}=(TP\times P-FP\times O)/\mathrm{n}=TP/\mathrm{n}\times P-FP/\mathrm{n}\times O$$

*(n=the total numbers of testing data=79325*0.3)*

Through the above equations, expected extra income of three models is calculated below:



Model	Expected Extra Income Ratio
Logistic Regression	0.531*P-0.051*O
Random Forest	0.53*P-0.052*O
KNN	0.49*P-0.092*O
Decision Tree	0.489*P-0.093*O

Based on the calculation, we can conclude using the overbooking strategy based on the cancellation prediction from Logistic Regression can give our client more extra income and the performance of Random Forest is also similar.

Deployment

1. How to deploy our model for business decision:

Integration with Property Management System(PMS): This predictive model will be integrated into this hotel's PMS. When a new reservation comes into record, the model will forecast the likelihood of cancellation.

Overbooking Decision: The hotel will make overbooking decisions according to the model's predictions. The default setting of this model is when the probability of cancellation is equal or larger than 50%, this reservation is likely to be canceled and overbooking is recommended to be implemented. This default benchmark could be adjusted accordingly.

2. Ethical considerations

Data Privacy: The hotel should respect data privacy law and protect guest information. The data should only be used for the purpose of managing reservations and cancellations.

Fairness: The hotel should guarantee that all data will be fairly used. The overbooking decision should only be based on model prediction rather than specific factors such as nationality, gender or age.

3. Risks and mitigation action

Overbooking risk: Although overbooking can optimize occupancy and maximize profits, there is a risk of overcommitting since the best model we chose still has false positives. It may cause guest dissatisfaction and affect the hotel brand or the need to relocate guests and raise compensation costs higher than the loss of leaving an empty room. Therefore, the hotel should carefully set overbooking limits initially and gradually increase them as the hotel gains confidence in the model's predictions. Moreover, the hotel can clearly communicate the overbooking policies to guests during the reservation process.

Correlation does not mean causality: Since our model is only a predictive model which can output whether the customer with certain features will cancel the booking, these features cannot help to decrease the cancellation rate. The hotel still needs to investigate the causality of cancellation.

Reference

Antonio, N., de 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

Cancellation trends cause headaches for Hotels. Duetto. (2016). https://www.duettocloud.com/library/cancellation-trends-cause-headaches-hotels