Predicting Booking Completion: Key Insights from Malaysian Travelers on British Airways*

Cristina Su Lam

October 2, 2024

Table of contents

1 Introduction **1**

2 Methods **2**

3 Results **3**
3.1 Model Selection, Diagnostics, Validation 3
3.2 Description of the Data 6

4 Discussion **8**
4.1 Final Model Interpretation 8
4.2 Limitations 8

5 Appendix **9**

References **9**

1 Introduction

The global aviation industry operates within a dynamic environment, necessitating a comprehensive understanding of the factors that influence booking completion rates among travelers. Hence our goal is to equip industry stakeholders with invaluable insights to inform strategic marketing initiatives and tailor offerings to meet customer expectations. This study focuses on travelers originating from Malaysia who consider British Airways as their preferred airline,

*Code and data supporting this analysis is available at: https://github.com/cristinaasu/BritishAirways_BookingAnalysis

presenting a unique opportunity to explore the intricacies of customer behavior within this specific context.

Previous research has highlighted the impact of factors like length of stay and flight duration on successful ticket bookings. Notably, a substantial proportion of tickets are booked for short flights, with just a one-day duration of stay at the destination (Eggermond, Schüssler, and Axhausen 2007). Additionally, individual studies have examined variables such as sales channel (Mohd Suki and Mohd Suki 2017), preference of in-flight meals (Lim and Lee 2019), and seat choices (Lim and Lee 2019), shedding light on their respective influences on booking completion. However, despite larger-scale investigations addressing the relationship between customer behavior and booking completion, research specific to the Malaysian context remains limited. Therefore, this study endeavors to fill this gap by incorporating all mentioned variables into a logistic regression model, with a particular emphasis on the local market.

2 Methods

The research utilizes a Kaggle dataset containing 50,000 observations of flight booking data from British Airways (Bisht 2024). We narrow our focus to 7,174 observations of travelers originating from Malaysia, ensuring a more targeted analysis. This dataset has 12 predictors, including key variables such as length of stay, flight duration, sales channel, in-flight meals, and seat choices, central to our investigation. To uncover the customer behavior factors influencing the binary outcome booking completion, we employ a Generalized Linear Model's logistic regression framework with a logit link function.

This statistical analysis begins with constructing a full model incorporating all variables of the dataset. Subsequently, variable selection techniques are applied to identify the most relevant predictors, such as stepwise AIC and BIC. While both methods involve iteratively adding/removing variables to minimize the AIC/BIC, stepwise BIC imposes a stronger penalty for additional parameters. Moreover, the LASSO method is used, penalizing the absolute size of regression coefficients to shrink less important predictors to zero. Model diagnostics are then conducted to assess the quality of each model, involving identifying outliers, influential points using DFBETAS, and assessing multicollinearity with VIF, where a value greater than 5 indicates multicollinearity among predictors. To validate the models, cross-validation is performed to evaluate prediction accuracy using MAE, alongside utilizing ROC curves with AUC value to assess discriminatory ability. These validation techniques ensure the robustness and generalizability of our models for accurate predictions on unseen data.

In the event of detecting outliers and influential points, only outliers will be removed due to their potential to bias parameter estimates. Influential points, although characterized by extreme values on predictor variables, will not be removed as they do not significantly impact the response variable. Following outliers removal, the model selection, diagnostics, and validation processes will be repeated using the cleaned dataset.

The primary objective is to identify the best-performing model with a lower MAE for accuracy and higher AUC for discrimination. Additionally, we need to consider the significance of our predictors of interest in the model. Upon selecting the final model, we conduct Exploratory Data Analysis (EDA) using histograms and contingency tables to enhance comprehension of underlying data patterns. Overall, this approach ensures a systematic and rigorous analysis of the association between customer behavior factors and ticket booking completion.

3 Results

3.1 Model Selection, Diagnostics, Validation

We initiated by fitting the full model, encompassing all the 12 variables from our dataset. Subsequently, we employed the variable selection techniques, LASSO method and stepwise AIC and BIC. Interestingly, the LASSO Method did not identify any variables for inclusion, likely due to the high penalty imposed, resulting in all coefficients being reduced to zero. Consequently, we proceeded with the models generated by stepwise AIC and BIC. Below are the variables selected by each method:

1. Stepwise AIC: "sales_channel", "length_of stay", "flight_duration", "num_of_passengers", "wants_extra_baggage", "wants_preferred_seat", "wants_in_flight_meals"
2. Stepwise BIC: "sales_channel", "length_of stay", "flight_duration", "wants_extra_baggage", "wants_preferred_seat", "wants_in_flight_meals"

Based on this, each model underwent thorough diagnostics and validation, which identified different numbers of influential points and outliers. As outlined in the methods section, we repeated the model selection, diagnostics, and validation processes using the cleaned dataset. Specifically, stepwise AIC and BIC procedures were re-ran with each cleaned dataset, yielding the same variables as above. Despite consistent variable selection, slight changes in model diagnostics and validation were observed. This iterative process enabled us to assess the impact of outlier removal on overall model performance, as summarized in Table 1.

Table 1: Comparison between different models

Model	Outliers	Influential Points	MAE	AUC	VIF
Stepwise AIC	13	1591	0.008	0.61	Predictors < 5
Stepwise AIC (cleaned data)	8	1570	0.006	0.61	Predictors < 5
Stepwise BIC	12	1351	0.008	0.60	Predictors < 5
Stepwise BIC (cleaned data)	7	1304	0.006	0.61	Predictors < 5

Drawing from the outcomes presented, the model derived from the Stepwise BIC approach with cleaned data is chosen. This decision is supported by several factors: the Stepwise BIC method penalizes complexity more heavily than AIC, prioritizing model simplicity and avoiding overfitting, which aligns with our research objective of identifying an accurate yet parsimonious model. The slightly lower MAE Figure 1 suggests better prediction accuracy compared to other models, while maintaining a comparable AUC Figure 2, indicating the model's ability to discriminate between completed and incomplete bookings. Moreover, as shown in Appendix Table 2, all predictors of interest exhibit strong statistical significance, with p-values less than 0.001, reinforcing the reliability and predictive power of the model.

```
[1] "sales_channel"         "length_of_stay"        "wants_preferred_seat"
[4] "wants_extra_baggage"   "wants_in_flight_meals" "flight_duration"
```

```
Logistic Regression Model

lrm(formula = booking_complete ~ ., data = cleaned_data_bic[,
    which(colnames(cleaned_data_bic) %in% c(newselect_var_bic,
        "booking_complete"))], model = TRUE, x = TRUE, y = TRUE)

                      Model Likelihood       Discrimination    Rank Discrim.
                            Ratio Test              Indexes          Indexes
Obs          7162    LR chi2     251.94       R2       0.048    C       0.609
 0           4706    d.f.             6      R2(6,7162)0.034    Dxy     0.218
 1           2456    Pr(> chi2) <0.0001    R2(6,4841.4)0.050    gamma   0.219
max |deriv| 6e-12                            Brier    0.218    tau-a   0.098

                     Coef    S.E.   Wald Z Pr(>|Z|)
Intercept            -0.4393 0.1174 -3.74  0.0002
sales_channel=Mobile -0.4332 0.0857 -5.05  <0.0001
length_of_stay       -0.0095 0.0010 -9.77  <0.0001
wants_extra_baggage   0.4158 0.0635  6.54  <0.0001
wants_preferred_seat  0.2057 0.0572  3.60  0.0003
wants_in_flight_meals 0.2258 0.0544  4.15  <0.0001
flight_duration      -0.0632 0.0151 -4.17  <0.0001


n=7162   Mean absolute error=0.006   Mean squared error=5e-05
0.9 Quantile of absolute error=0.012
```

During model diagnostics, although influential points and outliers were identified, their occurrence was less frequent compared to alternative models. In Figure 3, the length of stay showed

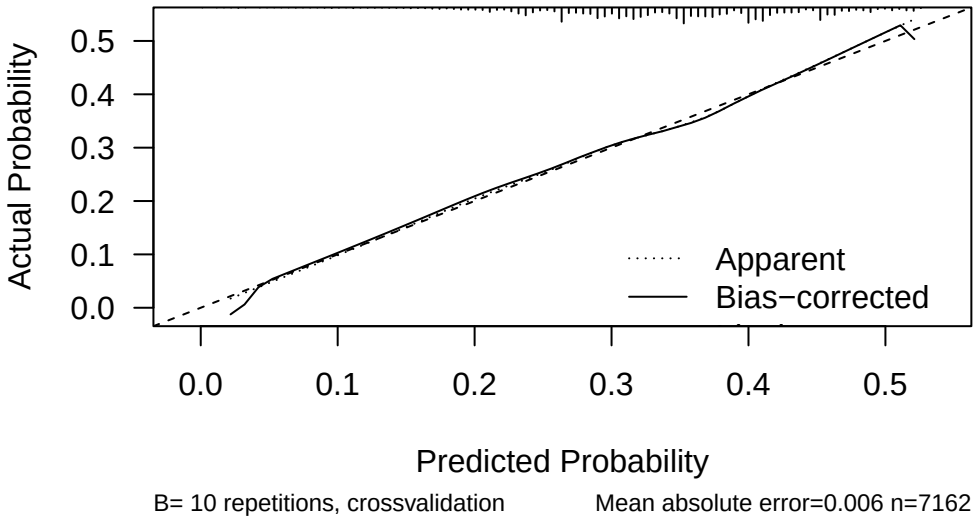

Figure 1: The Calibration Plot

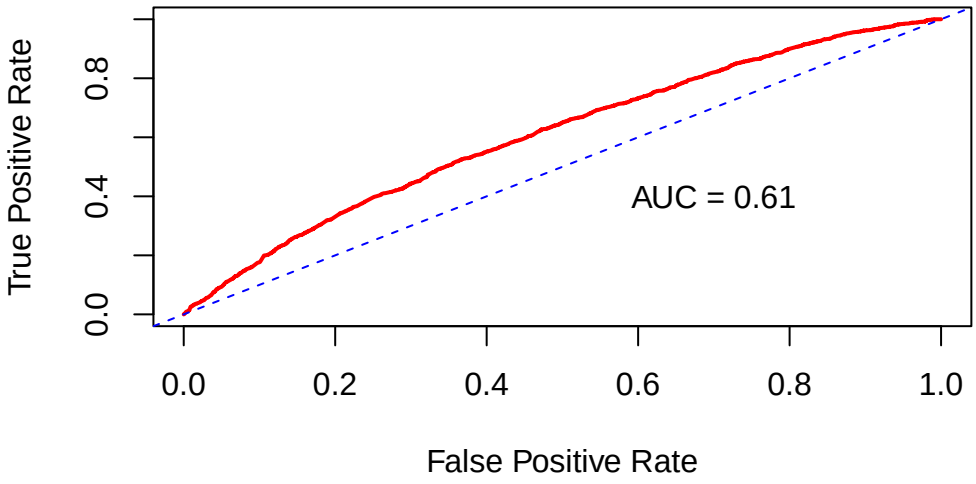

Figure 2: ROC Curve

an increase in influential points as the duration extended, possibly indicating special cases such as long vacations. In contrast, flight duration in Figure 4 exhibited fewer influential points affecting booking completion, suggesting it had less impact on the outcome. Additionally, Appendix Table 3 shows that all VIF values are under 5, confirming that multicollinearity is not a concern in our model. These results underscore the enhanced data quality and robustness achieved through some data cleaning procedures.

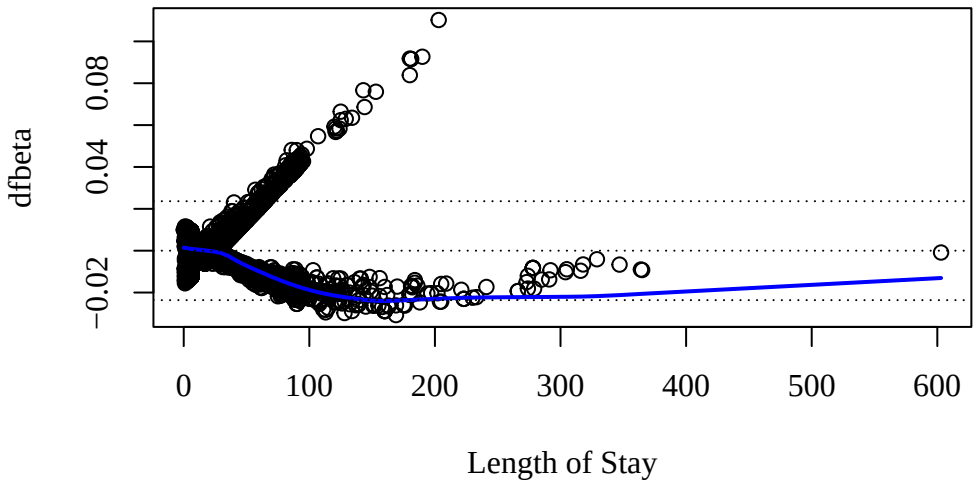

Figure 3: DFBETA vs. Length of Stay

3.2 Description of the Data

From our selected model, incorporating variables suggested by stepwise BIC from the cleaned dataset, we extracted significant insights into the distribution of customer behavior factors and their impact on booking completion. Examining the distributions depicted in Figure 5 and Appendix. Table 4 and Table 5, only 2456 bookings were completed, with the majority (2242) being made through the company's online website. A large proportion of passengers opted for extra baggage (1968), a smaller proportion chose preferred seating, and nearly half of the customers preferred to forgo in-flight meals. Surprisingly, the substantial proportion of incomplete bookings underscores the need for further investigation into the factors affecting this response variable, we expect that our model can make a meaningful contribution in this regard.

In conclusion, the Stepwise BIC approach with cleaned data, tailored to represent the general

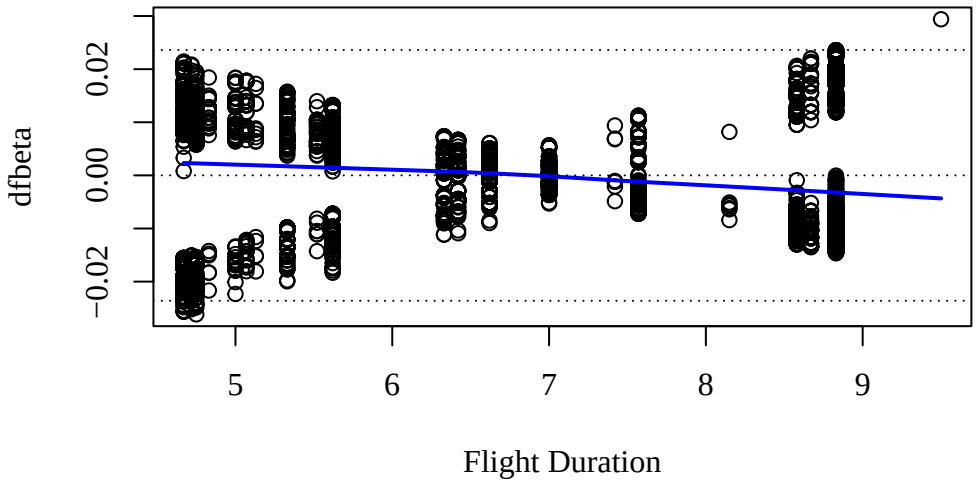

Figure 4: DFBETA vs. Flight Duration

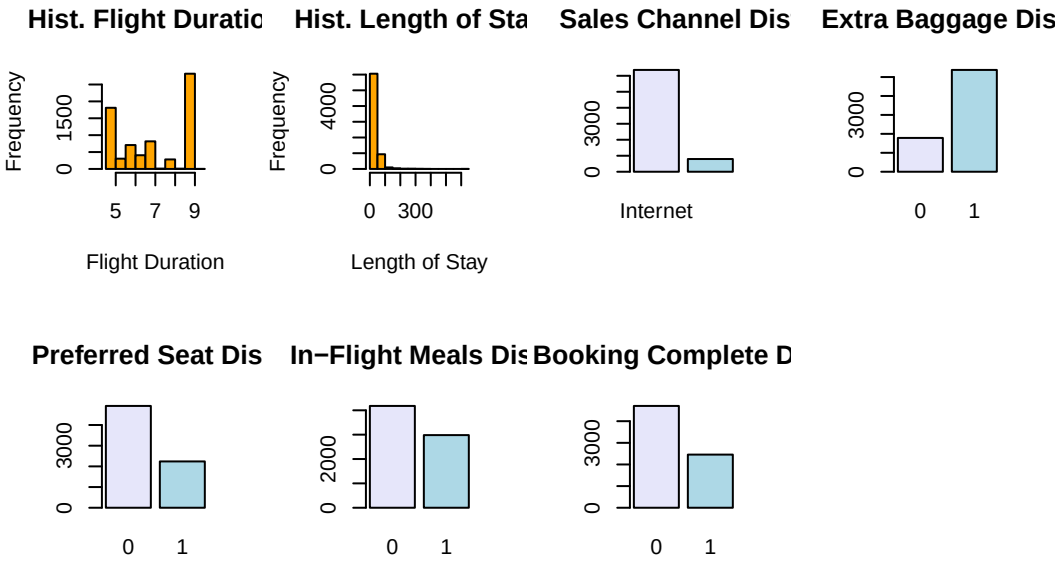

Figure 5: Histogram and Bar-plots of Variables

customer base of the company in Malaysia, achieves an optimal balance between predictive accuracy, model simplicity, and robustness. This renders it the preferred choice for examining the core factors influencing booking completion among travelers originating from Malaysia and considering British Airways as their airline of choice.

4 Discussion

4.1 Final Model Interpretation

The final model uncovers insightful relationships between customer behavior factors and booking completion. Significantly, the odds of booking completion for customers using the mobile sales channel are 1.54 times lower (e^0.433 = 1.54) compared to those using the internet sales channel. This suggests potential challenges in the mobile application booking process. Conversely, additional services, such as extra baggage, preferred seating, and in-flight meals, positively impacts booking completion rates (Appendix. Table 2)

Moreover, the statistically significant coefficients underscore the robustness of these relationships, indicating that the observed effects are not due to random variation but reflect genuine associations, and reinforces the reliability of our findings. Overall, the model addresses our research question by elucidating the key determinants impacting booking completion rates, providing actionable insights for industry stakeholders to optimize the booking process and enhance customer satisfaction.

4.2 Limitations

Acknowledging the stepwise BIC model with a cleaned dataset as our preferred choice for addressing the research question, it's crucial to note several limitations. While the AUC value is acceptable, a higher value closer to 1 would be preferable, improving the model's ability to distinguish class separation. Achieving this optimal value may require adjustments to the model architecture.

Despite efforts to remove outliers, a significant number of influential points and outliers persist in the dataset, potentially distorting the model's performance and predictions. Addressing them typically requires complex data preprocessing techniques and careful consideration of research's goal.

Therefore, while acknowledging these limitations, it's important to interpret the results of the final model cautiously and consider their potential implications for decision-making within the aviation industry. Future research could explore alternative strategies for handling influential points and outliers to enhance the model's predictive accuracy and robustness.

5 Appendix

Table 2: Summary Statistics Table for Stepwise BIC Model with Cleaned Dataset

Coefficients	Estimate	Std. Error	p-value
Intercept	-0.439	0.117	<0.001
Sales Channel/Mobile	-0.433	0.086	<0.001
Length of Stay	-0.009	0.001	<0.001
Flight Duration	-0.063	0.015	<0.001
Extra Baggage	0.416	0.064	<0.001
Preferred Seat	0.206	0.057	<0.001
In-flight Meals	0.226	0.054	<0.001

Table 3: VIF values for Stepwise BIC Model with Cleaned Dataset

Sales Channel	Length of Stay	Flight Duration	Extra Baggage	Preferred Seat	In-flight Meals
1.010675	1.049261	1.038505	1.091465	1.134532	1.13929

Table 4: Contingency Table for Sales Channel

	0	1	Total
Internet	4125	2242	6367
Mobile	581	214	795
Total	4706	2456	7162

Table 5: Contingency Table for Wants Extra Baggage

	0	1	Total
0	1298	488	1786
1	3408	1968	5376
Total	4706	2456	7162

References

Bisht, D. 2024. "British Airways Customer Bookings." https://www.kaggle.com/datasets/deepakb4/british-airways-customer-bookings.

Eggermond, M. van, N. Schüssler, and K. W. Axhausen. 2007. "Consumer Choice Behaviour and Strategies of Air Transportation Service Providers." ETH, Eidgenössische Technische Hochschule Zürich, IVT, Institut für Verkehrsplanung und Transportsysteme.

Lim, J., and H. C. Lee. 2019. "Comparisons of Service Quality Perceptions Between Full Service Carriers and Low Cost Carriers in Airline Travel." *Current Issues in Tourism* 23 (10): 1261–76. https://doi.org/10.1080/13683500.2019.1604638.

Mohd Suki, N., and N. Mohd Suki. 2017. "Flight Ticket Booking App on Mobile Devices: Examining the Determinants of Individual Intention to Use." *Journal of Air Transport Management* 62: 146–54. https://doi.org/10.1016/j.jairtraman.2017.04.003.