MSE 718

Group 5

Final Project

Flight Prices Prediction

Chenyu(Vivian) Liu 21143173

Sammi(Yi-Shan) Liu 21161990

Luna(Yijun) Lu 21169134

Christy(Jinning) Zhang 20894922

Names are sorted by the initial letter of the surname

1. Introduction

As global air travel expands, airlines face increasing complexity and competition, driving a need for deeper insights into airfare pricing. Prices vary widely due to dynamic strategies and factors such as travel dates, flight duration, and airline type. The business environment is increasingly characterized by unpredictability, complexity, and rapid changes due to new technologies and rising competition on the Internet (Narangajavana et al., 2014). This complexity motivates our project to explore advanced predictive techniques, specifically Bayesian methods, which integrate prior knowledge and observed data to handle uncertainty. The project aims to identify underlying patterns and interactions among key variables influencing airfare, providing more robust insights than traditional analyses and supporting strategic decision-making in the airline industry.

2. Methods

2.1 Data Description

The data for our research is sourced from the EaseMyTrip Flight Fare Details 2020 on Kaggle. The dataset contains flight fare information collected through web scraping from easemytrip.in for the period January 1, 2020, to February 29, 2020, with a total of 1,794,624 records. It includes fields such as flight operator, flight number, departure and arrival time, layovers, number of stops, and fare prices.

2.2 Data Processing

Based on the original variables, we created some new features to help our analysis:

- Total Minutes: Total duration of the flight in minutes.
- Distance: Distance between the departure and arrival locations in kilometres.
- IsWeekend: A binary feature indicating whether the departure date falls on a weekend (Friday, Saturday, or Sunday).
- If Holiday: A binary feature indicating whether the arrival date is a *holiday* [1].
- Is Low Cost: A binary variable identifying whether the airline is classified as a *low-cost carrier* [2].
- Departure Off Peak: A binary feature indicating whether the departure doesn't occur during peak hours (8am - 9pm).
- Arrival Off Peak: A binary feature indicating whether the arrival doesn't occur during peak hours.

2.3 Data Cleaning

Our dataset exhibited severe right-skewness in airfare distribution due to extremely high fares, violating assumptions of normality required by certain modeling methods. To mitigate potential bias, we employed the Interquartile Range (IQR) method to remove outliers and non-value data points, achieving a cleaner and more symmetric distribution for analysis.

2.4 Exploratory Data Analysis

2.4.1 Univariate analysis

- Total_Minutes: Right-skewed; 68.86% flights last 300–1,500 minutes, with 6.7% less than 3,000.
- Distance: Right-skewed; 70.94% flights are under 3,000 km, 4.66% exceed 10,000 km.
- Low_Cost_Count: 80% have no low-cost carriers; few have 1-3.
- Number-of-Stops: Most flights have 1 stop; fewer have 2; very few are nonstop or have 3 stops.
- If_Weekend: 33.9% depart on weekends; 66.1% on weekdays.
- If_Holiday: 5.03% depart on holidays.

- If_Low_Cost: 24.80% are low-cost; 75.20% are not.
- Departure_Off_Peak: 30.50% depart off-peak.
- Arrival_Off_Peak: 30.72% arrive off-peak.

2.4.2 Bivariate analysis

(a) Correlation between flight fares and predictors. This also supports our interaction terms:

- Categorical variables V.S. fare

  Nonstop flights show lower median fares but more high-priced outliers. Fares increase with stops; two-stop flights are the costliest. As low-cost carrier availability increases, median fares decrease, indicating price reductions due to low-cost segments.
- Numerical variables V.S. fare

  Both duration and distance show a positive relationship with fare separately, where longer flights tend or longer distances generally command higher ticket prices.
- Binary variables V.S. fare

  Weekend flights and off-peak departures show minimal impact on fares. Flights with low-cost carriers notably reduce median fares compared to traditional airlines. Arrivals during off-peak hours have a wider fare distribution and higher median, suggesting greater variability.

(b) Correlations among predictors:

To ensure predictor independence, we analyzed inter-predictor correlations. Based on our constructed *hit map* [4], we found that IsWeekend and ifHoliday,Is-Low-Cost and distance, Total-Minutes and Number.Of.Stops, Departure.Off.Peak and Arrival.Off.Peak, Total-Minutes and ifHoliday, Total-Minutes and IsWeekend, Departure.Off.Peak and Number.Of.Stops have more significant relations.

2.5 Related Research

Chen et al. (2015) highlight airfare volatility and recommend exploring multi-stop routes, individual flights, and airlines for deeper insights. Liu et al. (2017) developed the Adaptive Context-Aware Ensemble Regression (ACER) model, dynamically adapting to market changes using an ensemble of predictive techniques, including Bayesian methods. However, their Bayesian application was limited, indicating opportunities for deeper integration. Boruah et al. (2019) demonstrated Bayesian methods' effectiveness in managing uncertainty by employing Kalman filters to predict airfare based on historical data, further suggesting potential for advanced Bayesian approaches in airfare prediction.

2.6 Methodology

Our research contributed by explicitly integrating interaction terms, addressing Chen et al. (2015)'s call for exploring complex routes and Liu et al. (2017)'s focus on context-aware modeling. Additionally, inspired by Boruah et al. (2019), we advance Bayesian methods by employing a hierarchical Bayesian framework to better capture airfare volatility.

Our main research question is the influence of diverse variables on airline ticket pricing, specifically, how do interaction terms among factors such as dates, airline type and more impact pricing? Our project starts with Linear Regression to establish a comprehensive baseline of factors influencing airfare. Subsequently, we adopt the Hamiltonian Monte Carlo (HMC) for the Bayesian Model with all important factors. This aims to enhance our estimation accuracy.

The exploration of interaction terms through Linear Regression:

- IsWeekend:ifHoliday
  - Hypothesis: Weekend moderates holiday airfare effects. Following Koo & Mantin (2010), holiday weekends may intensify price dispersion due to higher leisure travel and varied pricing strategies.
- Is_Low_Cost:distance
  - Hypothesis: Airline type moderates the relationship between travel distance and airfare. Building on Wehner et al. (2018), differences in pricing strategies between low-cost and traditional carriers likely affect how these airlines price flights across varying distances.
- Total_Minutes:Number.Of.Stops
  - Hypothesis: Number of stops confounds total flight time and airfare relationships. Martínez-Val et al. (2012) demonstrate that stops significantly influence costs, implying an interdependent impact with flight duration.
- Departure.Off.Peak:Arrival.Off.Peak
  - Hypothesis: Off-peak arrival time moderates the relationship between off-peak departure time and airfare. Building on Smith and Johnson (2021), who highlight how departure and arrival times significantly influence airline pricing strategies, we posit that their combined effect uniquely impacts airfare.
- Total_Minutes:ifHoliday
  - Hypothesis: Holidays mediate the relationship between total flight time and airfare by influencing flight selections, indirectly affecting airfare. Wen and Yeh (2017) emphasize that understanding traveller preferences during holidays helps airlines optimize pricing and scheduling.
- Total_Minutes: IsWeekend
  - Hypothesis: The weekend indicator moderates the relationship between total flight time and airfare. Puller (2012) noted larger weekend discounts on business-heavy routes compared to leisure routes, suggesting weekend booking impacts airfare differently based on flight duration.
- Departure.Off.Peak:Number.Of.Stops
  - Hypothesis: Off-peak departure time acts as a collider between the number of stops and airfare. Escobari (2017) found systematically higher fares during peak times, indicating that departure timing significantly influences airfare decisions related to route complexity.

We utilized ANOVA to confirm the significance of these interactions, subsequently incorporating significant terms into our Linear and Bayesian models. Model validation employed R-squared and k-fold cross-validation.

Given the relatively high RMSE observed, we further explored four advanced models:

- Non-linear model - Capture the complex and often non-linear relationships inherent in the airfare pricing data
- Random forest - Avoid overfitting and handle complex datasets with numerous variables effectively, making it ideal for the multifaceted nature of airfare data.
- Feature engineering - Refine and create new predictors from existing data, potentially unveiling hidden patterns that more straightforward models might miss.
- Hierarchical Bayesian model - Inspired by the successful application of Bayesian methods with Kalman filter (Boruah et al., 2019), we propose a hierarchical Bayesian model to delve deeper into the data hierarchy.

3. Results

3.1 Linear Regression Model

$$Fare = B_0 + B_1\, Number\ of\ Stops + B_2\, Total\ Travel\ Minutes + B_3\, Distance + B_4\, Is\ Weekend + B_5\, Holiday\ Period + B_6 LowCost\ Airline + B_7\, Low\ Cost\ Airline\ Count + B_8\, Off\ Peak\ Departure + B_9 OffPeak\ Arrival + \varepsilon$$

All variables are significant as their p-values are smaller than 0.05[5].

3.2 Bayesian Regression Result

The Bayesian R-squared is 0.4617, indicating that about 46.17% of the variance in airfare is explained by the model[6]. Both the linear and Bayesian regression models exhibit similar R-squared values, with Bayesian regression only slightly outperforming linear regression. This suggests that the complexity of airfare pricing is not fully captured by traditional predictors alone. We further explore interaction terms.

3.3 Interaction Term

Model	P_value	Sum_of_Sq
IsWeekend:ifHoliday	0.05712 .	806583409
Is_Low_Cost:distance	2.2e-16 ***	7.3568e+10
Total_Minutes:Number.Of.Stops	2.2e-16 ***	1.7716e+10
Departure.Off.Peak:Arrival.Off.Peak	0.01037 *	1464187478
Total_Minutes:ifHoliday	0.1761	407926201
Total_Minutes:IsWeekend	0.2946	244793203
Departure.Off.Peak:Number.Of.Stops	0.4301	138755764

Figure 1. ANOVA results for all interaction term

Figure 1 indicates that only three interaction terms are statistically significant, supporting the validity of three specific hypotheses.

- Is_Low_Cost:distance - Low-cost carriers adopt distance-dependent pricing.
- Total_Minutes:Number.Of.Stops - Stops and flight duration jointly influence airfare.
- Departure.Off.Peak:Arrival.Off.Peak - Off-peak departure and arrival uniquely affect pricing.

3.4 Models with Interaction Terms

$$Fare = B_0 + B_1 \, Number\ of\ Stops + B_2 \, Total\,Travel\,Minutes + B_3 \, Distance + B_4 \, Is\,Weekend + B_5 \, Holiday\ Period + B_6 LowCost\ Airline + B_7 \, LowCost\ Airline\ Count + B_8 \, OffPeak\ Departure + B_9 OffPeak\ Arrival + \gamma_1 Is\_Low\_Cost{:}\, distance + \gamma_2 Total\_Minutes{:}\, Number.Of.Stops + \gamma_3 Departure.Off.Peak{:}\, Arrival.Off.Peak + \mathcal{E}$$

These terms were incorporated into Linear and Bayesian models validated through R-squared and k-fold cross-validation. Bayesian Regression (R^2=0.4708, RMSE=0.7276500) slightly outperformed Linear Regression (R^2=0.4705, RMSE=0.7276648)[7].

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: Fare ~ Number.Of.Stops + Total_Minutes + distance + IsWeekend + ifHoliday + Is_Low_Cost + Low_Cost_Co
unt + Departure.Off.Peak + Arrival.Off.Peak + Is_Low_Cost:distance + Total_Minutes:Number.Of.Stops + Departure.Of
f.Peak:Arrival.Off.Peak
##    Data: flight_data (Number of observations: 27448)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Regression Coefficients:
##                                    Estimate Est.Error l-95% CI u-95% CI Rhat
## Intercept                           -218.09    595.05 -1369.99   988.11 1.00
## Number.Of.Stops                     5512.46    426.83  4674.38  6327.75 1.00
## Total_Minutes                         11.21      0.50    10.22    12.18 1.00
## distance                               3.59      0.03     3.52     3.66 1.00
## IsWeekend                           1355.07    189.58   979.57  1723.00 1.00
## ifHoliday                          -1617.46    410.98 -2441.58  -812.83 1.00
## Is_Low_Cost                         6665.24   1106.00  4449.01  8821.79 1.00
## Low_Cost_Count                     -5096.71    511.12 -6074.01 -4104.26 1.00
## Departure.Off.Peak                  1243.39    228.58   785.48  1687.54 1.00
## Arrival.Off.Peak                    3638.85    234.31  3173.87  4089.42 1.00
## distance:Is_Low_Cost                  -2.05      0.11    -2.25    -1.83 1.00
## Number.Of.Stops:Total_Minutes         -3.64      0.32    -4.27    -3.01 1.00
## Departure.Off.Peak:Arrival.Off.Peak -1482.81    435.97 -2318.23  -593.69 1.00
##                                    Bulk_ESS Tail_ESS
## Intercept                              1607     2362
## Number.Of.Stops                        1761     2115
## Total_Minutes                          1796     2283
## distance                               4382     2922
## IsWeekend                              5831     2349
## ifHoliday                              5306     2882
## Is_Low_Cost                            1807     2653
## Low_Cost_Count                         1982     2682
## Departure.Off.Peak                     3895     3134
## Arrival.Off.Peak                       4074     3269
## distance:Is_Low_Cost                   4137     2991
## Number.Of.Stops:Total_Minutes          1674     2068
## Departure.Off.Peak:Arrival.Off.Peak    3587     3132
##
## Further Distributional Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma 14802.58     61.97 14678.79 14929.10 1.00     7048     2281
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Figure 2. Results for Bayesian with Interaction Term

Interpretation for Bayesian Model with interaction term

- Intercept (B_0): The expected base fare is -$218.89 units when all other predictors are held constant.
- Number of Stops (B_1): Each additional stop increases airfare by approximately $5512.46, holding other variables constant.
- Total Minutes (B_2): Every additional minute of flight time raises the fare by $5.86, holding other variables constant.

- Distance (B_3): Every additional kilometre increases airfare by approximately \$3.59, holding other variables constant.
- Is Weekend (B_4): Flights on weekends are higher by \$1355.07 than on weekdays, holding other variables constant.
- If Holiday (B_5): Flights during holidays are lower by \$1617.46 compared to non-holidays, holding other variables constant.
- Is Low Cost (B_6): Flying with low-cost carriers reduces the fare by \$6665.24 compared to traditional carriers, holding other variables constant.
- Low-Cost Count (B_7): An increase in the availability of low-cost carriers reduces the fare by \$5096.71, holding other variables constant.
- Departure Off Peak (B_8): Departing during off-peak hours increases the fare by \$1243.39, holding other variables constant.
- Arrival Off Peak (B_9): Arriving during off-peak hours increases the fare by \$3638.85, holding other variables constant.
- Is_Low_Cost:distance (γ_1): Each additional unit of distance, the fare for low-cost airlines decreases by approximately \$2.05 more than it would for traditional airlines, holding other variables constant.
- Total_Minutes:Number.Of.Stops (γ_2): Each additional minute and each additional stop, the fare decreases by \$3.64, holding other variables constant.
- Departure.Off.Peak:Arrival.Off.Peak (γ_3): Flights both departing and arriving during off-peak hours are priced approximately \$1482.39 lower than those that do not, holding other variables constant.

3.5 Advanced Model

Due to high RMSE in our initial models, we further evaluated advanced approaches[8]:

```
##                       Model      RMSE  Rsquared
## 1          non-linear model 0.6874839 0.5274200
## 2             random forest 0.5338309 0.7152616
## 3       feature engineering 0.7246163 0.4749353
## 4 hierarchical bayesian model 0.7276561 0.4710922
```

Figure 3. Advanced Model CV Result

- Non-linear model: Captured complex non-linear airfare relationships, modestly improving predictability.
- Random Forest: Achieved the best performance (RMSE=0.5338, R^2=0.7153) through effective handling of data complexity and robustness to overfitting.
- Feature Engineering: Provided limited improvements, suggesting minimal gains from additional derived features.
- Hierarchical Bayesian model: Inspired by Boruah et al. (2019), it slightly improved performance (R^2=0.4780) by capturing hierarchical data but was less effective than ensemble methods like Random Forest.

4. Discuss

4.1 Conclusion

Our primary goal was to leverage Bayesian methodologies for airfare prediction, building upon prior studies emphasizing their effectiveness. While our Basic and Hierarchical Bayesian models enhanced insights into complex interactions, the Random Forest model exhibited superior predictive accuracy, likely due to its ability to manage complex data and resistance to overfitting. Our research contributes by explicitly integrating interaction terms within Bayesian frameworks, enhancing context-awareness, and advancing hierarchical Bayesian modeling techniques based on existing literature.

4.2 Implication

This project highlights how predictive modelling, incorporating Bayesian methods and other sophisticated techniques, can enhance insights and practical pricing strategies. For airlines, such models improve market forecasting and revenue management, while consumers benefit from fairer, more tailored pricing, enhancing travel satisfaction. Ultimately, these insights support strategic airline decisions and promote consumer welfare.

4.3 Limitation

A key limitation of our study was the unavailability of critical data, such as seat availability at purchase, limiting our model's ability to fully capture the factors influencing airfare prices. This lack of data restricted the exploration of hierarchical Bayesian and other models' full predictive potential. Future research could integrate multiple predictive techniques, including Bayesian methods, to better address the complexities inherent in airfare pricing datasets.

5. References

Boruah, A., Baruah, K., Das, B., Das, M. J., & Gohain, N. B. (2019). A Bayesian approach for flight fare prediction based on Kalman filter. In C. R. Panigrahi, B. Pati, P. Mohapatra, K. C. Pati, & R. Buyya (Eds.), *Progress in advanced computing and intelligent engineering* (pp. 191–203). Springer. https://doi.org/10.1007/978-981-13-0224-4_18.

Chen, Y., Cao, J., Feng, S., & Tan, Y. (2015). An ensemble learning based approach for building airfare forecast service. In *2015 IEEE International Conference on Big Data (Big Data)* (pp. 964–969). IEEE. https://doi.org/10.1109/BigData.2015.7363846.

Escobari, D. (2017, June 22). Airport, airline and departure time choice and substitution patterns: An empirical analysis. *Transportation Research: PartA Policy and Practice 103, 198–210.* https://doi.org/10.1016/j.tra.2017.05.034

Liu, T., Cao, J., Tan, Y., & Xiao, Q. (2017). ACER: An adaptive context-aware ensemble regression model for airfare price prediction. *2017 International Conference on Progress in Informatics and Computing (PIC)*, Nanjing, China, 312–317. https://doi.org/10.1109/PIC.2017.8359563.

Mantin, B., & Koo, B. (2010). Weekend effect in airfare pricing. *Journal of Air Transport Management, 16*(1), 48–50. https://doi.org/10.1016/j.jairtraman.2009.07.002.

Martínez-Val, R., Perez, E., Cuerno, C., & Palacin, J. F. (2012). Cost-range trade-off of intermediate stop operations of long-range transport airplanes. *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, *227*(2), 394–404. https://doi.org/10.1177/0954410011429766

Narangajavana, Y., Garrigos-Simon, F. J., García, J. S., & Forgas-Coll, S. (2014). Prices, prices and prices: A study in the airline sector. *Tourism Management, 41*, 28–42. https://doi.org/10.1016/j.tourman.2013.08.008.

Puller, S. L., & Taylor, L. M. (2012, October 3). Price discrimination by day-of-week of purchase: Evidence from the U.S. Airline Industry. *Journal of Economic Behavior and Organization 84(3), 801–812.* https://doi.org/10.1016/j.jebo.2012.09.022

Wehner, C., López-Bonilla, J. M., López-Bonilla, L. M., & Santos, J. A. C. (2018). State of the art of pricing policy in air transportation: Network carriers vs. low-cost airlines. *Tourism & Management Studies*, 14(3), 32–40. https://doi.org/10.18089/tms.2018.14303.

Wen, C.H., & Yeh, Y. (2017). Modeling air travelers' choice of flight departure and return dates on long holiday weekends. *Journal of Air Transport Management, 65*, 220–225. https://doi.org/10.1016/j.jairtraman.2017.06.016

6. Appendix

1) We set flight departures on 2020-01-01(New Year's Day), 2020-01-1(Makar Sankranti), 2020-01-26 (Chinese New Year), 2020-02-21 (Maha Shivaratri) as Holiday flights.
2) The low cost flight carriers include: AirAsia, Air Connect, Air India Express, AirArabia, Eurowings, FlyDubai, Flybe, GoAir, Hahn Air, Indigo, Jazeera Airways, Jeju Air, Jetstar Airways, Jetstar Asia, Jetstar Pacific, National Air Services, Onur Air, Pegasus Airlines, S7 Airlines, Scoot, SpiceJet, Thai Lion Air, Thai Smile, Trujet, WestJet Airlines.
3) IQR method:

First, we computed the first (Q1) and third (Q3) quartiles of the Fare distribution and used these to derive the IQR = Q3 − Q1. Next, we established lower and upper bounds at Q1-1.5*IQR, Q3+1.5*IQR, respectively. Any data points falling outside these bounds were considered outliers and removed from the dataset. This approach effectively minimized the influence of extreme values while preserving the majority of observations, thus providing a more balanced and reliable foundation for subsequent analyses.

4) Variable heat map:

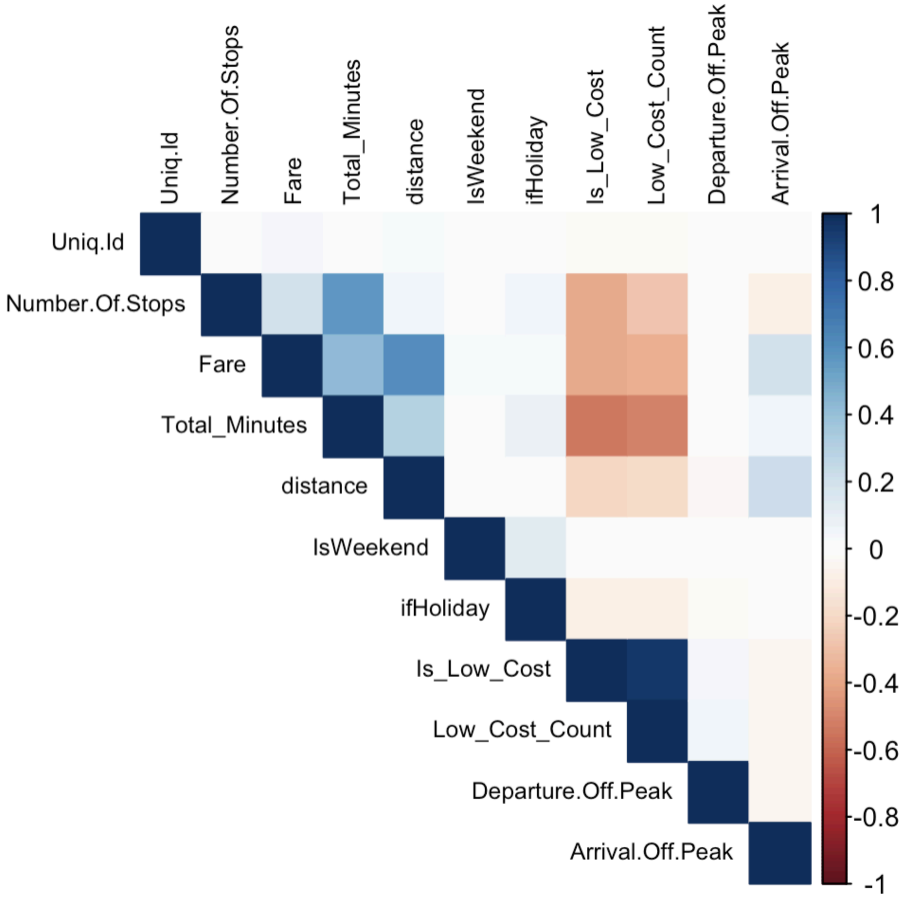

Figure 4.

5) Linear regression result:

```
##
## Call:
## lm(formula = Fare ~ Number.Of.Stops + Total_Minutes + distance +
##     IsWeekend + ifHoliday + Is_Low_Cost + Low_Cost_Count + Departure.Off.Peak +
##     Arrival.Off.Peak, data = flight_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -58312  -8929  -2390   4275  71496
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        6.610e+03  3.365e+02  19.642  < 2e-16 ***
## Number.Of.Stops    1.101e+03  2.142e+02   5.139 2.79e-07 ***
## Total_Minutes      5.858e+00  2.215e-01  26.454  < 2e-16 ***
## distance           3.415e+00  3.258e-02 104.804  < 2e-16 ***
## IsWeekend          1.351e+03  1.919e+02   7.037 2.01e-12 ***
## ifHoliday         -1.657e+03  4.178e+02  -3.967 7.29e-05 ***
## Is_Low_Cost       -4.486e+03  8.700e+02  -5.156 2.54e-07 ***
## Low_Cost_Count    -1.960e+03  4.291e+02  -4.567 4.96e-06 ***
## Departure.Off.Peak 6.226e+02  1.965e+02   3.168  0.00154 **
## Arrival.Off.Peak   3.167e+03  2.015e+02  15.718  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14930 on 27438 degrees of freedom
## Multiple R-squared:  0.4617, Adjusted R-squared:  0.4615
## F-statistic:  2615 on 9 and 27438 DF,  p-value: < 2.2e-16
```

Figure 5. Linear Regression Result

Intercept (B_0): The model intercept was estimated at 6,610, which suggests that the base price of a ticket, absent the influence of other variables, starts at approximately $6,610.

Significant Predictors:

- Number of Stops (B_1): Each additional stop is associated with an increase of approximately $1,101 in the fare, indicating that direct flights are typically cheaper than those with multiple stops, holding other variables constant.
- Total Minutes (B_2): Every additional minute of travel time is associated with a fare increase of about $5.86, holding other variables constant.
- Distance (B_3): Each additional kilometer is associated with an increase of approximately $3.42 in the fare, holding other variables constant.
- Is Weekend (B_4): Traveling on weekends is associated with an increase of approximately $1,351 in fare compared to weekdays, holding other variables constant.
- If Holiday (B_5): Traveling during a holiday period is associated with a decrease of approximately $1,657 in fare, holding other variables constant.
- Is Low Cost (B_6): Booking with a low-cost airline is associated with a decrease of approximately $4,486 in fare compared to other airlines, holding other variables constant.
- Low Cost Count (B_7): Each additional low-cost airline operating on the route is associated with a decrease of approximately $1,960 in fare, holding other variables constant.
- Departure Off Peak (B_8): Departing during off-peak hours is associated with an increase of approximately $626 in fare, holding other variables constant.

- Arrival Off Peak (B_9): Arriving during off-peak hours is associated with an increase of approximately $3,167 in fare, holding other variables constant.

The Adjusted R-squared is 0.4615. This means that about 46.15% of the variance in airfare is explained by the model. The P-value of the model is < 2.2e-16, which suggests the model is significant.

6) Bayesian regression with interaction terms result:

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: Fare ~ Number.Of.Stops + Total_Minutes + distance + IsWeekend + ifHoliday + Is_Low_Cost + Low_Cost_Co
unt + Departure.Off.Peak + Arrival.Off.Peak
##    Data: flight_data (Number of observations: 27448)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Regression Coefficients:
##                    Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept           6611.38    338.58  5938.45  7284.36 1.00     5183     3322
## Number.Of.Stops     1093.81    212.59   671.24  1507.45 1.00     3177     3365
## Total_Minutes          5.86      0.22     5.43     6.30 1.00     3619     3004
## distance               3.41      0.03     3.35     3.48 1.00     4713     2836
## IsWeekend           1351.55    188.54   980.24  1719.22 1.00     3784     3239
## ifHoliday          -1655.24    413.53 -2432.24  -848.47 1.00     3707     2738
## Is_Low_Cost        -4512.59    834.49 -6118.60 -2819.18 1.00     2372     2681
## Low_Cost_Count     -1946.86    413.55 -2791.68 -1157.12 1.00     2353     2667
## Departure.Off.Peak   621.76    198.27   247.74  1017.94 1.00     4231     2752
## Arrival.Off.Peak    3164.32    196.03  2783.53  3540.85 1.00     4365     2354
##
## Further Distributional Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma 14930.67     65.95 14803.43 15063.72 1.00     7000     2908
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Figure 6. Bayesian with Importance Predictors

7) Linear regression with interaction terms result:

```
##
## Call:
## lm(formula = Fare ~ Number.Of.Stops + Total_Minutes + distance +
##     IsWeekend + ifHoliday + Is_Low_Cost + Low_Cost_Count + Departure.Off.Peak +
##     Arrival.Off.Peak + Is_Low_Cost:distance + Total_Minutes:Number.Of.Stops +
##     Departure.Off.Peak:Arrival.Off.Peak, data = flight_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -61972  -8648  -2540   4088  73270
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        -2.200e+02  5.874e+02  -0.375 0.707977
## Number.Of.Stops                     5.511e+03  4.231e+02  13.025  < 2e-16 ***
## Total_Minutes                       1.122e+01  5.029e-01  22.306  < 2e-16 ***
## distance                            3.589e+00  3.373e-02 106.419  < 2e-16 ***
## IsWeekend                           1.356e+03  1.903e+02   7.128 1.04e-12 ***
## ifHoliday                          -1.622e+03  4.144e+02  -3.915 9.07e-05 ***
## Is_Low_Cost                         6.657e+03  1.081e+03   6.159 7.41e-10 ***
## Low_Cost_Count                     -5.092e+03  5.037e+02 -10.109  < 2e-16 ***
## Departure.Off.Peak                  1.241e+03  2.327e+02   5.331 9.82e-08 ***
## Arrival.Off.Peak                    3.637e+03  2.362e+02  15.396  < 2e-16 ***
## distance:Is_Low_Cost               -2.049e+00  1.044e-01 -19.621  < 2e-16 ***
## Number.Of.Stops:Total_Minutes      -3.642e+00  3.219e-01 -11.314  < 2e-16 ***
## Departure.Off.Peak:Arrival.Off.Peak -1.477e+03  4.320e+02  -3.419 0.000628 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14800 on 27435 degrees of freedom
## Multiple R-squared:  0.4708, Adjusted R-squared:  0.4706
## F-statistic:  2034 on 12 and 27435 DF,  p-value: < 2.2e-16
```

Figure 7. Linear Regression with Interactions

8) Model performance plot (cross validation)

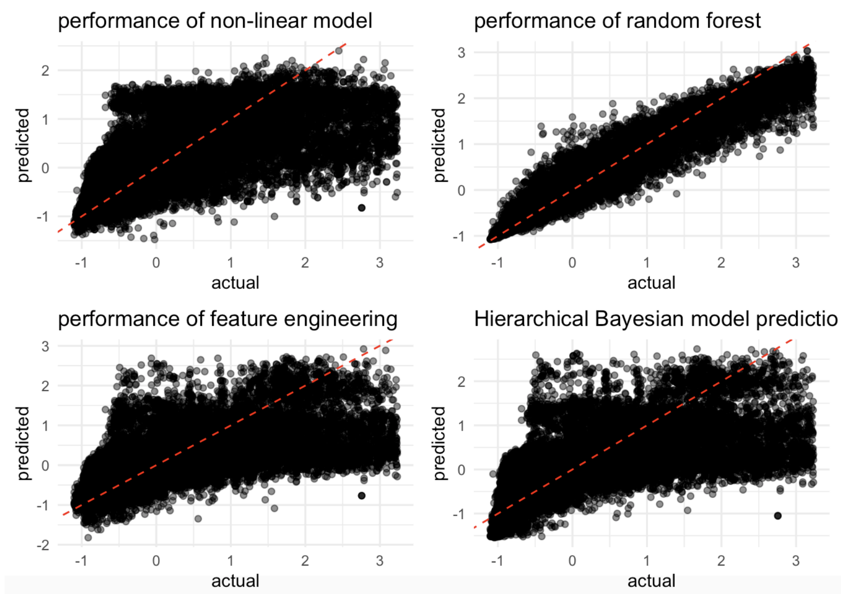

Figure 8. Model Performance