Airline Passenger Satisfaction: A Business Intelligence Approach

Team 57: Shawn Dai, Diane Guan, Taehui Kim, William Li, Sarvesh Vishnu Muhundu Parthiban

Business Understanding

Definition and Motivation

The airline industry has always been characterized by fierce competition and a relentless pursuit of passenger satisfaction. Airlines recognize that passenger experience is not just a reflection of service quality but also a significant driver of customer loyalty and, ultimately, profitability. In this Industry, optimizing passenger satisfaction is not merely a matter of customer service; it's a strategic Imperative.

Airline passengers today expect not only sage and punctual flights but also an exceptional overall experience. Factors such as in-flight services, seat comfort, cleanliness, entertainment options, and even online booking ease have all become integral to a passenger's perception of satisfaction. Airlines, in response to these heightened expectations, are continuously seeking ways to improve passenger satisfaction, streamline their operations, and gain a competitive edge in a crowded market.

Airlines operate in a dynamic and competitive environment, where success depends on a multitude of factors. Beyond the traditional operational challenges, airlines must also grapple with fluctuating fuel prices, regulatory requirements, and emerging trends, such as environmental sustainability. In this landscape, the ability to understand, predict, and influence satisfaction becomes a pivotal strategic advantage.

To enhance the ability to forecast, comprehend, and oversee the level of contentment among airline consumers, it is imperative to address a crucial inquiry in the realm of business: Given the different features of satisfaction rate contributors, how accurately can we predict the customer’s satisfaction rate? What are the principal determinants that serve as indicators of passenger satisfaction rates pertaining to airline services? Our goal is to improve the satisfaction rate of customers.

Data Mining Business Solutions

In an effort to identify the factors that contribute to passenger satisfaction, the airline industry has turned to data mining as a powerful tool. The dataset provided in the Kaggle repository offers a wealth of information, including passenger demographics, flight attributes, and satisfaction ratings. For this case, a predictive model and clustering method specifically k-prototypes are deployed for answering the business questions.

- What are the factors that contribute to passenger satisfaction
- Which passengers are more likely to be dissatisfied with their flight experiences?

The utilization of predictive models and clustering techniques in the field of machine learning presents a versatile methodology for comprehending and forecasting customer satisfaction levels within the domain of airline services. Predictive models inherently employ the analysis of the survey dataset in order to make predictions about future events. Within the framework of our business inquiry, individuals can receive training on diverse aspects pertaining to airline services in order to forecast the probability of customer satisfaction. This enables airline businesses to assess their ability to predict customer satisfaction based on specific features, facilitating proactive actions to improve the customer experience.

In contrast, clustering methods are utilized to partition customers into discrete groups according to their responses and preferences. Through this process, it is possible to uncover concealed patterns and connections within the dataset. As an illustration, a particular cluster could comprise of passengers who place a high emphasis on the availability of inflight entertainment options and the comfort of their seats, whereas another cluster might prioritize punctuality and the quality of the food served during the flight. The comprehension of these clusters has the potential to provide insights into the primary factors that influence satisfaction levels among various customer segments.

Data Understanding & Preparation

Variable Descriptions

Categorical Data:

- Gender: Indicates the gender of passengers.
- Customer Type: Differentiates between loyal and disloyal customers.
- Type of Travel: Specifies the purpose of the flight.
- Class: Denotes the travel class on the plane.

Ordinal Data:

- Satisfaction & Satisfaction Level: The Target variable satisfaction indicating the level of satisfaction typically ranges from 1-5 or 0-5.

- Similarly, ordered Survey data from 0-5 includes: Inflight_Wifi_Services, Ease of Online Booking, Gate Location, Food & Drink, Online Booking, Seat Comfort, Inflight Entertainment, On-board services, Leg room services, Baggage Boarding, Checking Service, Inflight Service, Cleanliness.

Numerical Data:

- Age: Age of the passengers.
- Flight Distance: Distance of the flight in miles.
- Departure Delay in Minutes: Duration of departure delay.
- Arrival Delay in Minutes: Duration of arrival delay.

In this analysis, we explore the factors that influence customer satisfaction in the airline industry. The first two columns in the dataset are index columns and do not contribute to the classification task. Hence, they have been excluded from the dataset. There exists a total of 310 instances of missing values within the Arrival Delay in Minutes column. To address this issue, the median value was employed as a method to impute the missing data points. In the dataset, certain features possess categorical characteristics, despite being classified as integer type. Consequently, it was necessary to convert these columns into the appropriate factor type/dummy variables.

Post Data Preparation, Data Exploration, Correlation analysis and further research our preliminary findings are as follows:

- Customer Satisfaction: 56.7% of passengers are neutral or dissatisfied, while 43.3% are satisfied.
- Customer Type: (81.7% Customers) Loyal Passengers have a significantly higher satisfaction rate (47.73%) than Disloyal customers (23.67%).
- Age: Passengers aged between 45 and 60 have the highest satisfaction rate. With an average passenger's age being 39 Years.
- Type of Travel: Business Travelers (Making up 68.9% of passengers))have a higher satisfaction rate (58.26%) than personal travelers.
- Class: Business Class passengers have the highest satisfaction rate (69.43%), followed by eco plus class (24.61%) and eco class (18.62%).
- Flight Distance: Passengers traveling over 1000 miles have the highest satisfaction rate (61.92%)
- Service Quality: Passengers rating inflight services at 4 & 5 have the highest satisfaction rate. Similar patterns of increasing satisfaction with service quality are observed across online booking, seat comfort, inflight entertainment, and other services.

Initial analysis suggests that customer satisfaction is influenced by a variety of factors, including type, age, type of travel, class, flight distance, and service quality. Arrival and departure delays also play a role in customer satisfaction.

Modeling

Core Task/ Predictive Classification Model: We evaluated three predictive models - logistic regression, random forest, and Lasso - using k-fold cross validation to test accuracy. Logistic regression estimates the likelihood of a binary event (Customer satisfaction or neutral/dissatisfaction) by examining the correlation between several attributes and generating a probability score. Logistic regression coefficients show how each factor affects the result, its significance, and directionality, revealing the factors that most affect customer satisfaction. Logistic regression might struggle to capture complex non-linear correlation in data, which random forest can by capturing complex non-linear interactions, potentially improving customer satisfaction prediction, especially when attributes and satisfaction are complex.This may reveal different traits' relative influence.

Finally, the Lasso regression algorithm is a linear regression model that adds a penalty term to the cost function to encourage sparse solutions where some coefficients are forced to be zero. This can help to improve the interpretability of the model and reduce overfitting.

Clustering method: K - Prototypes; The K- prototypes algorithm is a clustering methodology specifically developed to handle datasets that encompass a blend of categorical and numerical attributes. Given that our dataset contains categorical, ordinal and numerical, the K-prototypes algorithm is appropriate.

The utilization of clustering techniques in K-Prototypes enables the identification of discrete passenger segments by grouping individuals with similar characteristics. Each individual cluster may have unique characteristics that can impact overall satisfaction. After clustering, centroids can be analyzed to better understand each group's unique traits. This analysis reveals the main characteristics of satisfaction versus dissatisfaction clusters.

Evaluation

Prediction Models:

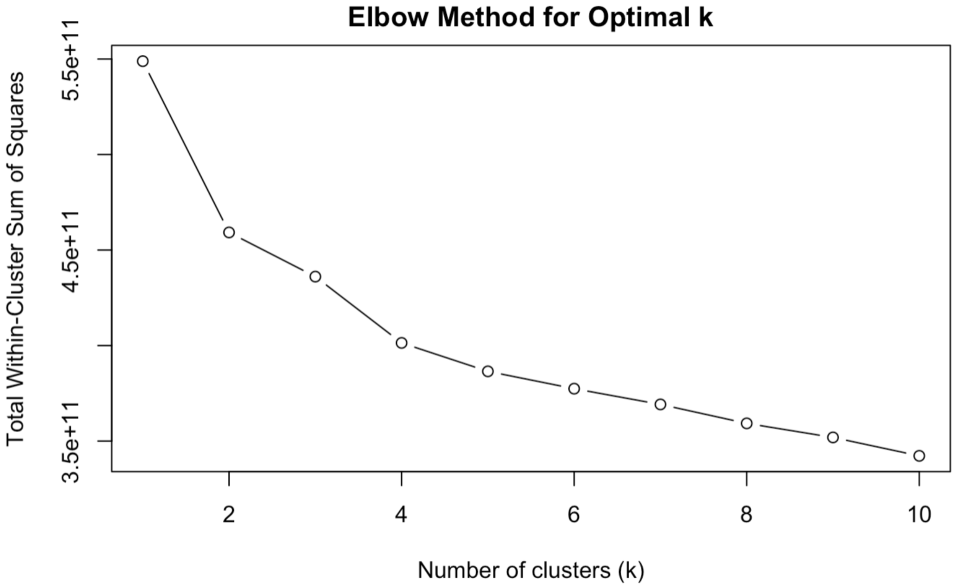
The first model that we tried is simple logistic regression. We used all of the features, used accuracy as the benchmark, and ran a 10 fold cross validation. The result is an accuracy value of 0.9339583. According to the p-values of the coefficient, the most statistically significant features are: the type of travel, class, arrival time convenient, ease of online booking, baggage handling, checkin service, inflight service, cleanliness, and arrival/departure delay. Most of the levels of these features are significant at the 99% level. An interesting note is that there is a warning message, fitted probabilities numerically 0 or 1 occurred on three of the levels, which are checkin service level 5, inflight service level 5, and cleanliness level 5, which means that all of the customers who rated these levels were satisfied.

The second model that we tried is random forest. We used all of the features, used accuracy as the benchmark, we only ran a 5 fold cross validation with ntree of 300 due to computational limitations. The best result is an accuracy value of 0.9640726 when mtry is 40.

The third model that we tried is Lasso. We used all of the features, used mean squared error as the benchmark, and ran a 10 fold cross validation to find the best lambda value, which is 0.0002554023, and the r-squared value is 0.5506358. This small lambda value means that the model will only benefit slightly from regularization, and the primary characteristics and relationships in your data are captured without applying a heavy penalty to the coefficients. The r squared value means that approximately 55.06% of the variability in the dependent variable can be explained by the independent variables in the model. According to the coefficients of the

LASSO model, the most important features are customer type, type of travel, class, and online boarding.

Clustering Model:



We use an elbow plot to determine the optimal number of clusters, and choose 4 clusters based on the significant drop in cluster sum of square value.

Cluster 1 consists of 26265 customers that travel mainly for business purposes, prefer business class, average age of 42.57, which is the oldest out of the four clusters, and have an average travel distance of 2768.95 miles, which is the longest out of the four clusters. There are more females than males in this cluster, most of the services such as inflight wifi, seat comfort, baggage handling, etc., have a rating of 4 out of 5. However, the inflight wifi service and ease of online booking are rated 2. This cluster is mostly satisfied.

Cluster 2 consists of 25655 customers that travel mainly for business purposes, but still prefer economic class, average age of 38.43, and have an average travel distance of 616.59 miles, which is the shortest out of the four clusters. There are more males than females in this cluster, their rate most of the services a 4, but they rated inflight wifi, departure/arrival time convenience and ease of online booking a 2, and they are mostly neutral or dissatisfied.

Cluster 3 consists of 27201 customers that travel mainly for personal purposes, prefer business class, average age of 40.04, and have an average travel distance of 951.83 miles. There are more males than females in this group, and their ratings are unbalanced. Most of the services have a rating of 5, indicating an excellent experience. However, the inflight wifi service, departure/arrival time convenience, and ease of online booking are rated 1. Despite the polarized rating, this group is still mostly satisfied.

Cluster 4 consists of 24783 customers that travel mainly for personal purposes, prefer economic class, average age of 37.4, which is the youngest out of the four clusters, and have an average travel distance of 725.24 miles. There are more females than males in this group. They rated most of the services a 3 or 4, however, they are still mostly neutral or dissatisfied.

Deployment

The results of the data mining project in the airline industry have the potential to significantly impact decision-making and customer satisfaction. To deploy these results effectively, consider the following steps:

- Integration with Business Operations: Integrate the predictive model and clustering insights into the airline's existing operations. Ensure that the models are seamlessly connected to customer feedback channels, reservation systems, and other relevant data sources .
- Establish a feedback loop that continuously updates the models with new data. Customer feedback and satisfaction ratings change over time, and it's crucial to adapt to evolving passenger preferences and market dynamics.

- Develop interactive data visualization tools that allow airline staff, from management to front-line employees, to easily access and interpret the results .Visualizations can provide actionable insights at a glance. To make informed decisions based on the insights generated.

Additionally, we must also consider issues regarding development and certain ethical considerations that we must consider.

- Privacy and Data Security: We must ensure that the customer data is handled with utmost care and its compliance with changing data privacy regulations. Anonymize and protect sensitive customer information.
- Model drift, continual improvement to be made to the model to adapt to the changing customer behavior.
- Fairness, Transparency: We must further ensure that models do not discriminate against any particular group of passengers based on sensitive attributes like gender, age, or race.

To ensure that the processes surrounding deployment remain fair, accurate, and up-to-date. Regular audits can be conducted. This can further help address any biases or inaccuracies promptly. Maintain open lines of communication with passengers, explaining how their data is used to enhance airline experience. Implement stringent data protection measures to safeguard customer data. Regularly update security protocols to counter emerging threats. We must also be prepared for the model predictions to not align with actual customer experiences and for that case have strategies in place to manage customer dissatisfaction and service recovery. Finally legal compliance to stay updated with relevant data protection and consumer privacy laws, follow responsible AI and ethical guidelines provided by organizations such as the Partnership on AI to endure the ethical use of AI and data mining.

Bibliography:

1. Brown, B., & Smith, J. (2017). Data Mining and Machine Learning: Fundamental Concepts and Algorithms, Morgan Kaufmann.
2. Witten, I.HFrank, E., & Hall, M.A. (2016). Data Mining: Practical Machine Learning Tools and Techniques.
3. Hand, D.J., Mannila, H., & Smyth, P. (2001). Principles of Data Mining. The MIT Press.
4. Shmueli, G., Patel, N.R., & Bruce, P.C. (2019). Data Mining for Business Analytics: Concepts, Techniques, and Applications.
5. Doganis, R. (2006) The Airline Business in the 21st Century. Routledge.

Individual Contributions:

Shawn Dai: Data Preparation, Modeling

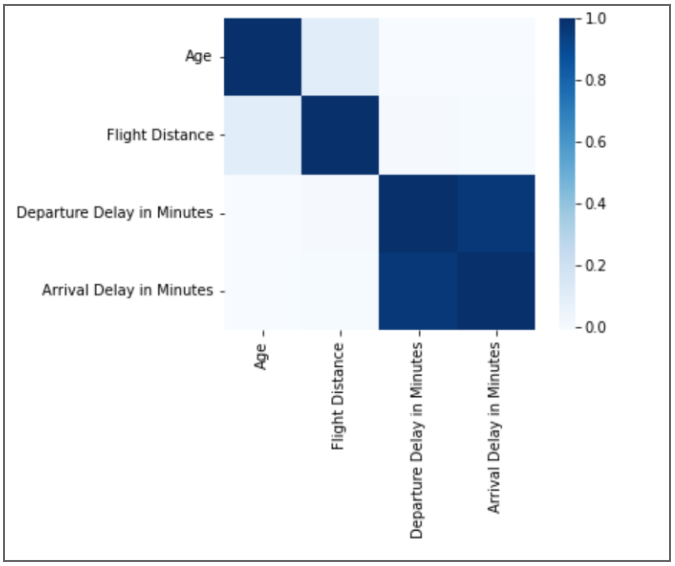
Diane Guan: Data Understanding, Data Preparation

Taehui Kim: Business Understanding, Data Understanding

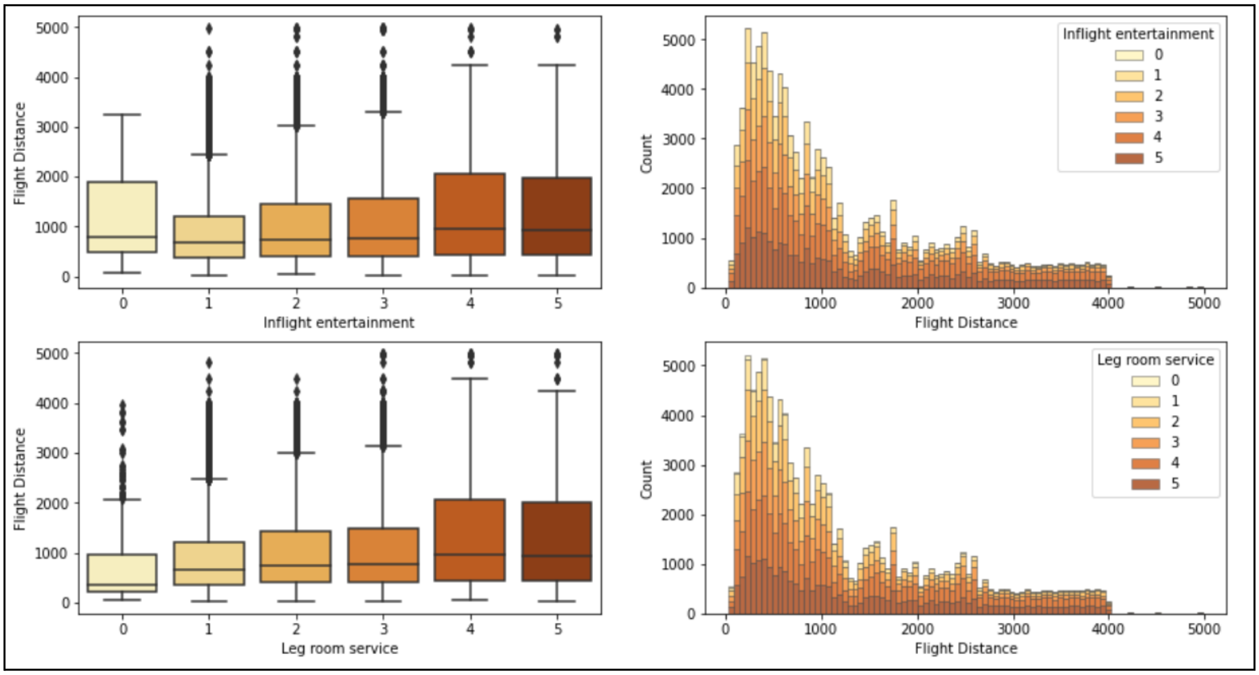
William Li: Modeling, Deployment

Sarvesh Vishnu Muhundu Parthiban: Business Understanding, Deployment

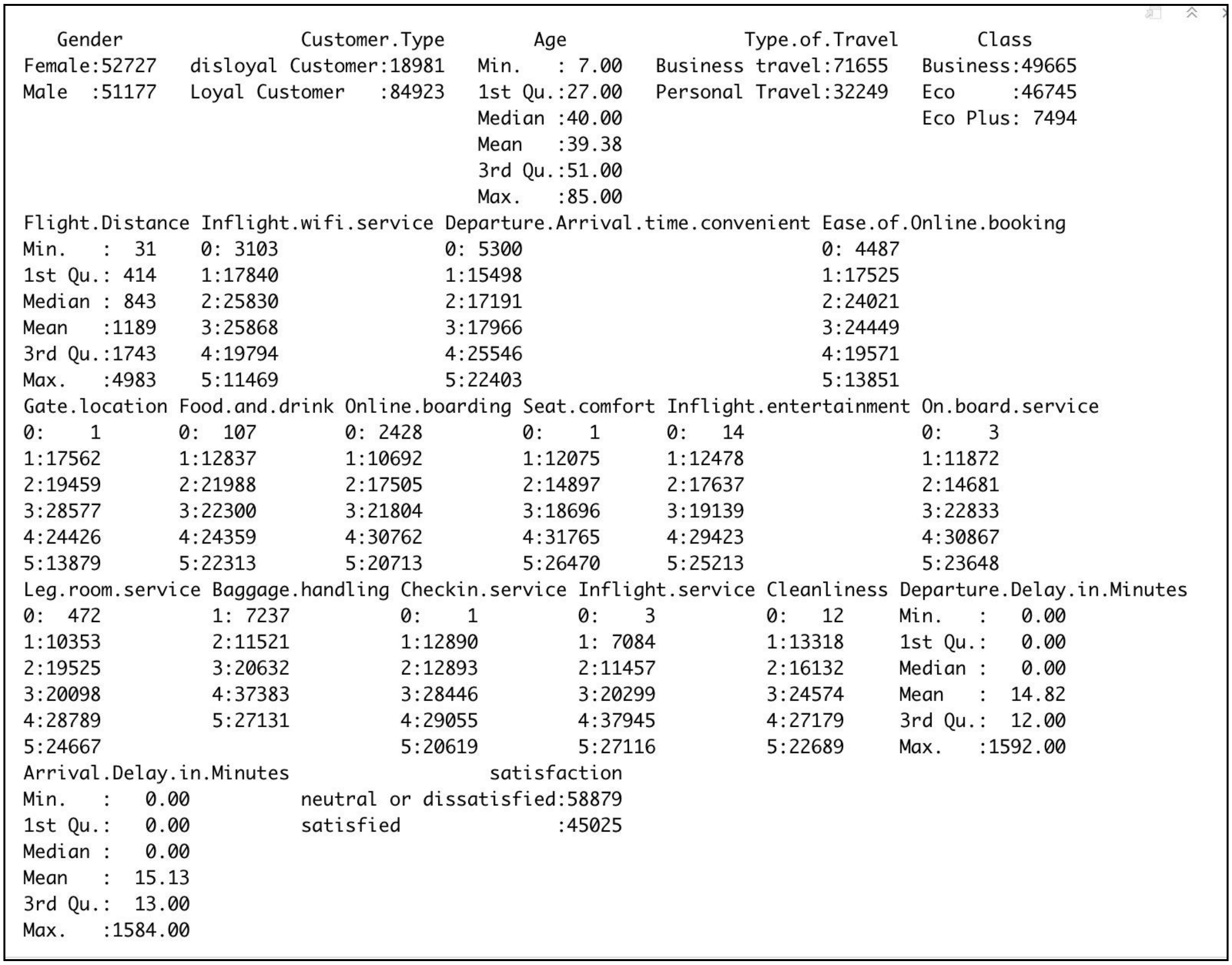
Appendix:



EDA; Corr-Matrix

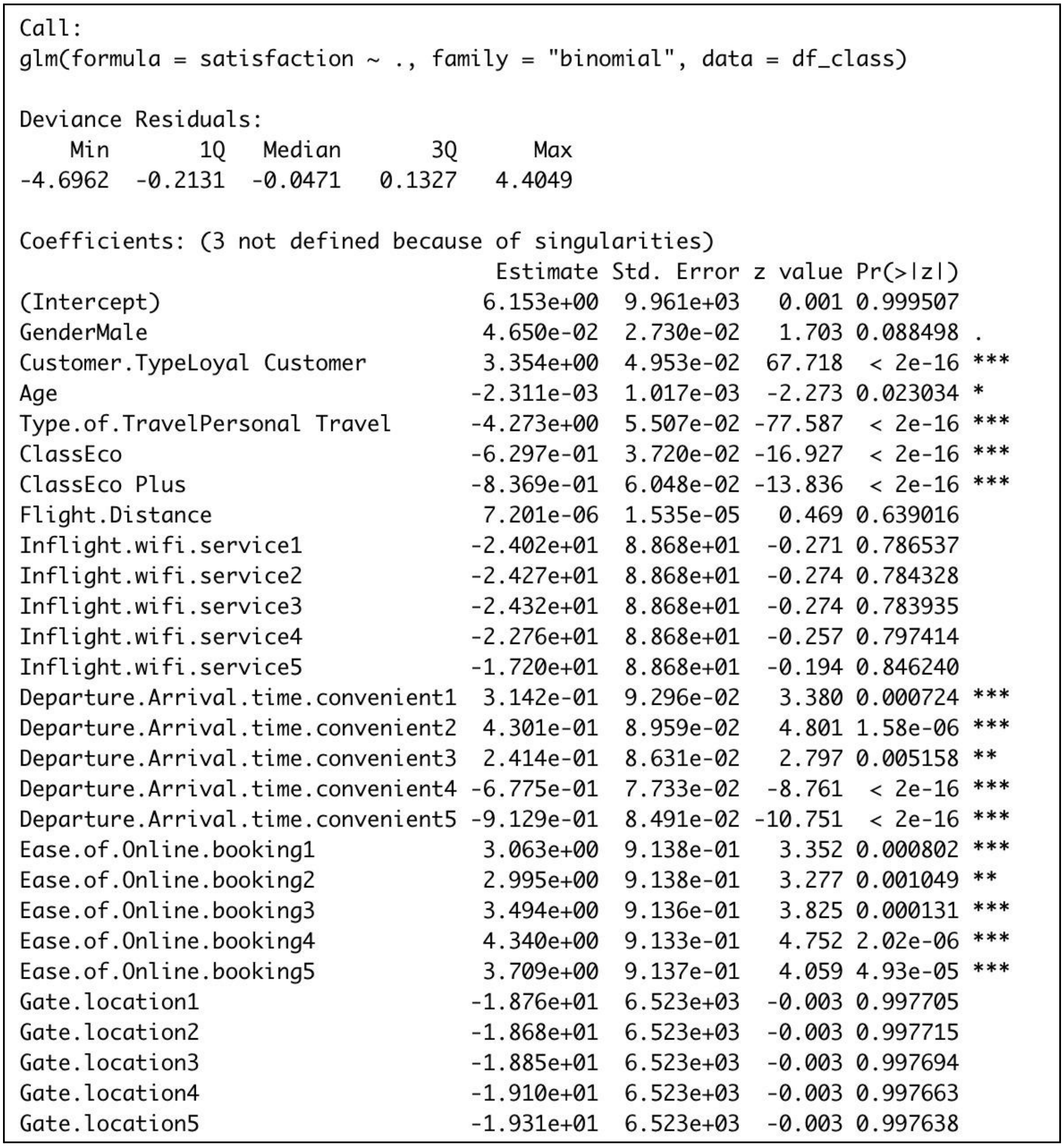


Data Summary

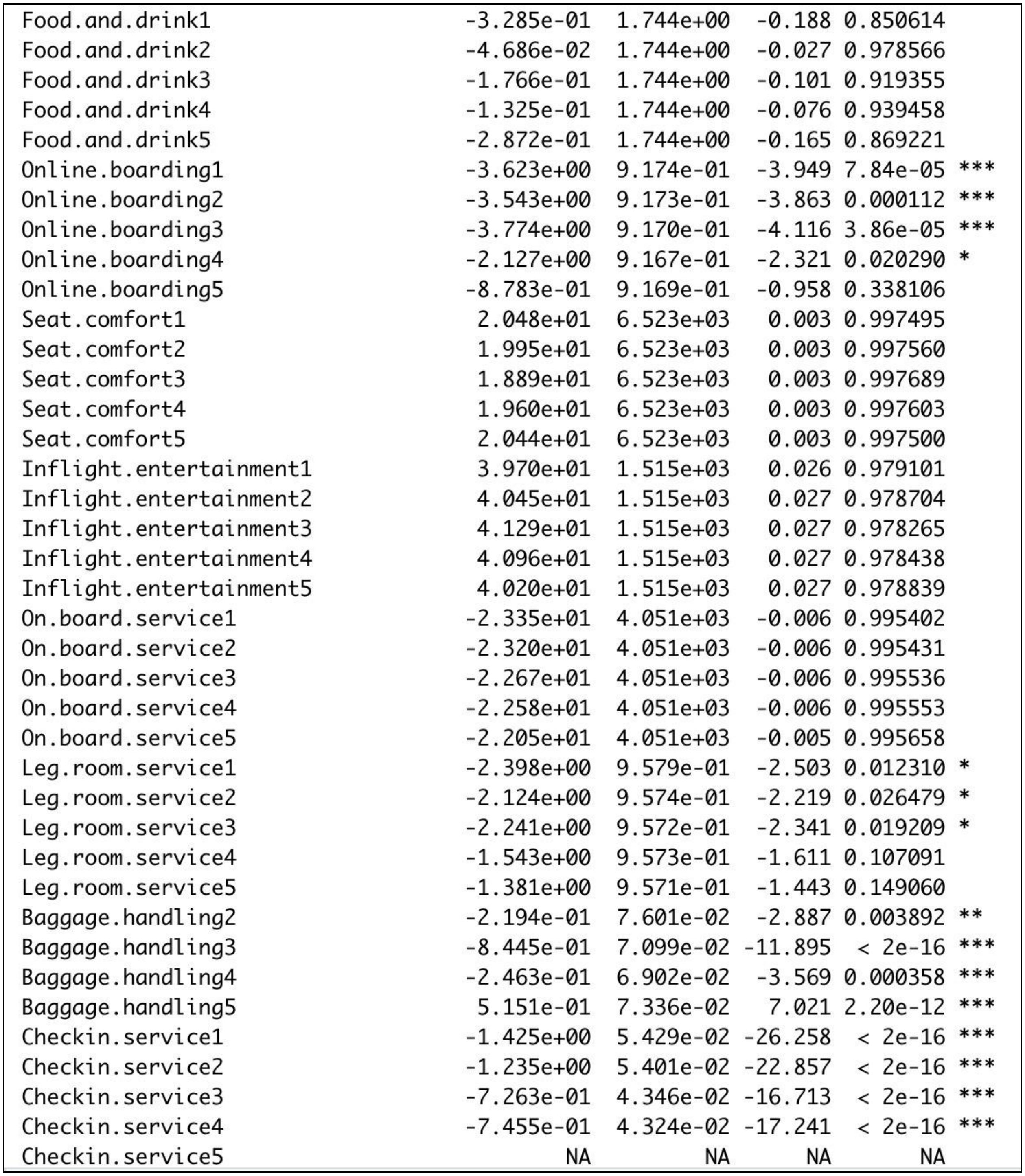


```
   Gender                Customer.Type        Age              Type.of.Travel       Class
Female:52727   disloyal Customer:18981   Min.   : 7.00   Business travel:71655   Business:49665
Male  :51177   Loyal Customer   :84923   1st Qu.:27.00   Personal Travel:32249   Eco     :46745
                                         Median :40.00                           Eco Plus: 7494
                                         Mean   :39.38
                                         3rd Qu.:51.00
                                         Max.   :85.00
Flight.Distance Inflight.wifi.service Departure.Arrival.time.convenient Ease.of.Online.booking
Min.   :  31    0: 3103               0: 5300                           0: 4487
1st Qu.: 414    1:17840               1:15498                           1:17525
Median : 843    2:25830               2:17191                           2:24021
Mean   :1189    3:25868               3:17966                           3:24449
3rd Qu.:1743    4:19794               4:25546                           4:19571
Max.   :4983    5:11469               5:22403                           5:13851
Gate.location Food.and.drink Online.boarding Seat.comfort Inflight.entertainment On.board.service
0:    1       0:  107        0: 2428         0:    1      0:   14                0:    3
1:17562       1:12837        1:10692         1:12075      1:12478                1:11872
2:19459       2:21988        2:17505         2:14897      2:17637                2:14681
3:28577       3:22300        3:21804         3:18696      3:19139                3:22833
4:24426       4:24359        4:30762         4:31765      4:29423                4:30867
5:13879       5:22313        5:20713         5:26470      5:25213                5:23648
Leg.room.service Baggage.handling Checkin.service Inflight.service Cleanliness Departure.Delay.in.Minutes
0:  472          1: 7237          0:    1         0:    3         0:   12      Min.   :   0.00
1:10353          2:11521          1:12890         1: 7084         1:13318      1st Qu.:   0.00
2:19525          3:20632          2:12893         2:11457         2:16132      Median :   0.00
3:20098          4:37383          3:28446         3:20299         3:24574      Mean   :  14.82
4:28789          5:27131          4:29055         4:37945         4:27179      3rd Qu.:  12.00
5:24667                           5:20619         5:27116         5:22689      Max.   :1592.00
Arrival.Delay.in.Minutes                  satisfaction
Min.   :   0.00          neutral or dissatisfied:58879
1st Qu.:   0.00          satisfied              :45025
Median :   0.00
Mean   :  15.13
3rd Qu.:  13.00
Max.   :1584.00
```

Logistic Regression

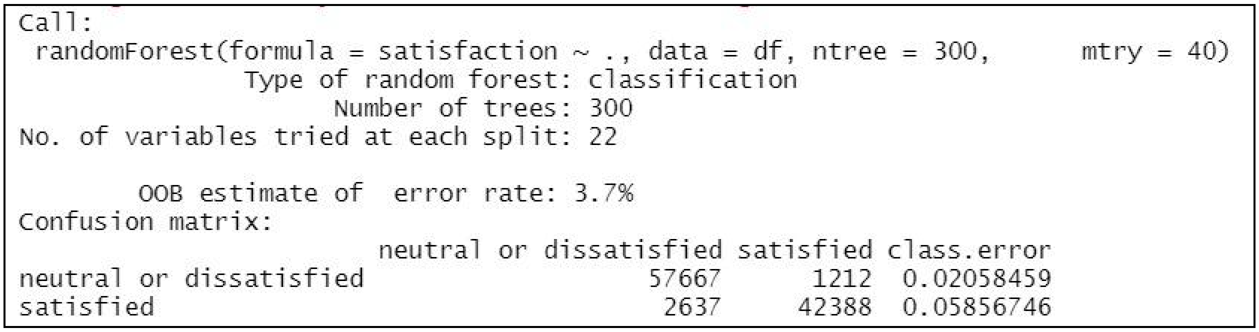


```
Call:
glm(formula = satisfaction ~ ., family = "binomial", data = df_class)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6962  -0.2131  -0.0471   0.1327   4.4049

Coefficients: (3 not defined because of singularities)
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                        6.153e+00  9.961e+03   0.001 0.999507
GenderMale                         4.650e-02  2.730e-02   1.703 0.088498 .
Customer.TypeLoyal Customer        3.354e+00  4.953e-02  67.718  < 2e-16 ***
Age                               -2.311e-03  1.017e-03  -2.273 0.023034 *
Type.of.TravelPersonal Travel     -4.273e+00  5.507e-02 -77.587  < 2e-16 ***
ClassEco                          -6.297e-01  3.720e-02 -16.927  < 2e-16 ***
ClassEco Plus                     -8.369e-01  6.048e-02 -13.836  < 2e-16 ***
Flight.Distance                    7.201e-06  1.535e-05   0.469 0.639016
Inflight.wifi.service1            -2.402e+01  8.868e+01  -0.271 0.786537
Inflight.wifi.service2            -2.427e+01  8.868e+01  -0.274 0.784328
Inflight.wifi.service3            -2.432e+01  8.868e+01  -0.274 0.783935
Inflight.wifi.service4            -2.276e+01  8.868e+01  -0.257 0.797414
Inflight.wifi.service5            -1.720e+01  8.868e+01  -0.194 0.846240
Departure.Arrival.time.convenient1  3.142e-01  9.296e-02   3.380 0.000724 ***
Departure.Arrival.time.convenient2  4.301e-01  8.959e-02   4.801 1.58e-06 ***
Departure.Arrival.time.convenient3  2.414e-01  8.631e-02   2.797 0.005158 **
Departure.Arrival.time.convenient4 -6.775e-01  7.733e-02  -8.761  < 2e-16 ***
Departure.Arrival.time.convenient5 -9.129e-01  8.491e-02 -10.751  < 2e-16 ***
Ease.of.Online.booking1            3.063e+00  9.138e-01   3.352 0.000802 ***
Ease.of.Online.booking2            2.995e+00  9.138e-01   3.277 0.001049 **
Ease.of.Online.booking3            3.494e+00  9.136e-01   3.825 0.000131 ***
Ease.of.Online.booking4            4.340e+00  9.133e-01   4.752 2.02e-06 ***
Ease.of.Online.booking5            3.709e+00  9.137e-01   4.059 4.93e-05 ***
Gate.location1                    -1.876e+01  6.523e+03  -0.003 0.997705
Gate.location2                    -1.868e+01  6.523e+03  -0.003 0.997715
Gate.location3                    -1.885e+01  6.523e+03  -0.003 0.997694
Gate.location4                    -1.910e+01  6.523e+03  -0.003 0.997663
Gate.location5                    -1.931e+01  6.523e+03  -0.003 0.997638
```



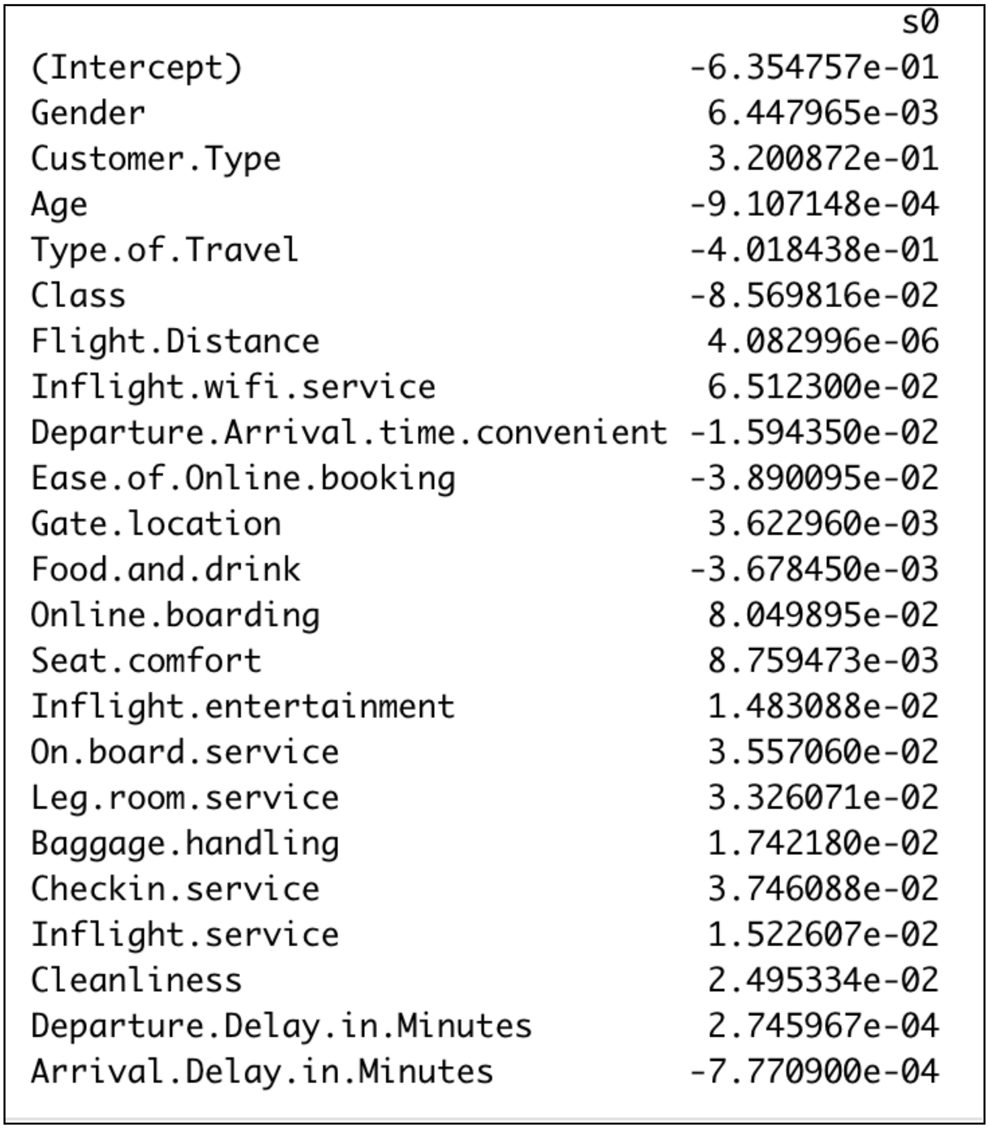
```
Food.and.drink1              -3.285e-01  1.744e+00  -0.188 0.850614    
Food.and.drink2              -4.686e-02  1.744e+00  -0.027 0.978566    
Food.and.drink3              -1.766e-01  1.744e+00  -0.101 0.919355    
Food.and.drink4              -1.325e-01  1.744e+00  -0.076 0.939458    
Food.and.drink5              -2.872e-01  1.744e+00  -0.165 0.869221    
Online.boarding1             -3.623e+00  9.174e-01  -3.949 7.84e-05 ***
Online.boarding2             -3.543e+00  9.173e-01  -3.863 0.000112 ***
Online.boarding3             -3.774e+00  9.170e-01  -4.116 3.86e-05 ***
Online.boarding4             -2.127e+00  9.167e-01  -2.321 0.020290 *  
Online.boarding5             -8.783e-01  9.169e-01  -0.958 0.338106    
Seat.comfort1                 2.048e+01  6.523e+03   0.003 0.997495    
Seat.comfort2                 1.995e+01  6.523e+03   0.003 0.997560    
Seat.comfort3                 1.889e+01  6.523e+03   0.003 0.997689    
Seat.comfort4                 1.960e+01  6.523e+03   0.003 0.997603    
Seat.comfort5                 2.044e+01  6.523e+03   0.003 0.997500    
Inflight.entertainment1       3.970e+01  1.515e+03   0.026 0.979101    
Inflight.entertainment2       4.045e+01  1.515e+03   0.027 0.978704    
Inflight.entertainment3       4.129e+01  1.515e+03   0.027 0.978265    
Inflight.entertainment4       4.096e+01  1.515e+03   0.027 0.978438    
Inflight.entertainment5       4.020e+01  1.515e+03   0.027 0.978839    
On.board.service1            -2.335e+01  4.051e+03  -0.006 0.995402    
On.board.service2            -2.320e+01  4.051e+03  -0.006 0.995431    
On.board.service3            -2.267e+01  4.051e+03  -0.006 0.995536    
On.board.service4            -2.258e+01  4.051e+03  -0.006 0.995553    
On.board.service5            -2.205e+01  4.051e+03  -0.005 0.995658    
Leg.room.service1            -2.398e+00  9.579e-01  -2.503 0.012310 *  
Leg.room.service2            -2.124e+00  9.574e-01  -2.219 0.026479 *  
Leg.room.service3            -2.241e+00  9.572e-01  -2.341 0.019209 *  
Leg.room.service4            -1.543e+00  9.573e-01  -1.611 0.107091    
Leg.room.service5            -1.381e+00  9.571e-01  -1.443 0.149060    
Baggage.handling2            -2.194e-01  7.601e-02  -2.887 0.003892 ** 
Baggage.handling3            -8.445e-01  7.099e-02 -11.895  < 2e-16 ***
Baggage.handling4            -2.463e-01  6.902e-02  -3.569 0.000358 ***
Baggage.handling5             5.151e-01  7.336e-02   7.021 2.20e-12 ***
Checkin.service1             -1.425e+00  5.429e-02 -26.258  < 2e-16 ***
Checkin.service2             -1.235e+00  5.401e-02 -22.857  < 2e-16 ***
Checkin.service3             -7.263e-01  4.346e-02 -16.713  < 2e-16 ***
Checkin.service4             -7.455e-01  4.324e-02 -17.241  < 2e-16 ***
Checkin.service5                     NA         NA      NA       NA    
```

Random Forest



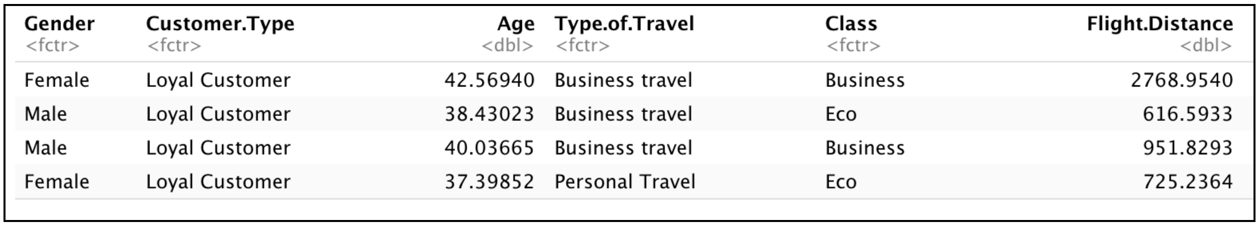
```
Call:
 randomForest(formula = satisfaction ~ ., data = df, ntree = 300,      mtry = 40)
               Type of random forest: classification
                     Number of trees: 300
No. of variables tried at each split: 22

        OOB estimate of  error rate: 3.7%
Confusion matrix:
                        neutral or dissatisfied satisfied class.error
neutral or dissatisfied                   57667      1212  0.02058459
satisfied                                  2637     42388  0.05856746
```

LASSO

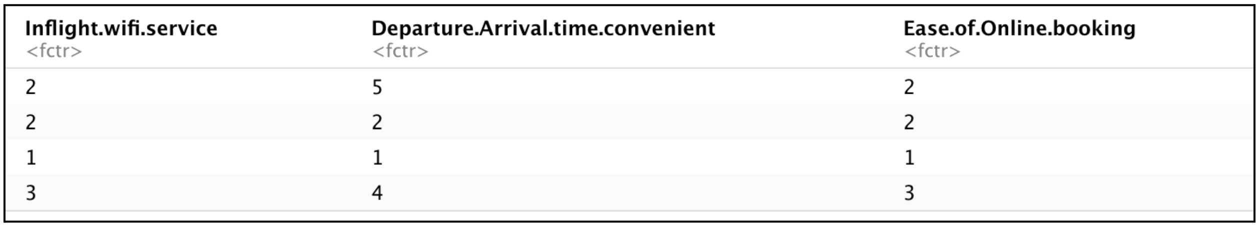


```
                                              s0
(Intercept)                        -6.354757e-01
Gender                              6.447965e-03
Customer.Type                       3.200872e-01
Age                                -9.107148e-04
Type.of.Travel                     -4.018438e-01
Class                              -8.569816e-02
Flight.Distance                     4.082996e-06
Inflight.wifi.service               6.512300e-02
Departure.Arrival.time.convenient -1.594350e-02
Ease.of.Online.booking             -3.890095e-02
Gate.location                       3.622960e-03
Food.and.drink                     -3.678450e-03
Online.boarding                     8.049895e-02
Seat.comfort                        8.759473e-03
Inflight.entertainment              1.483088e-02
On.board.service                    3.557060e-02
Leg.room.service                    3.326071e-02
Baggage.handling                    1.742180e-02
Checkin.service                     3.746088e-02
Inflight.service                    1.522607e-02
Cleanliness                         2.495334e-02
Departure.Delay.in.Minutes          2.745967e-04
Arrival.Delay.in.Minutes           -7.770900e-04
```

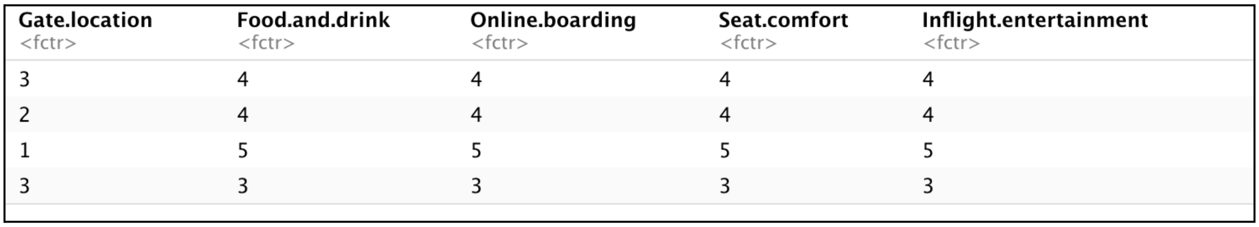
K-Proto



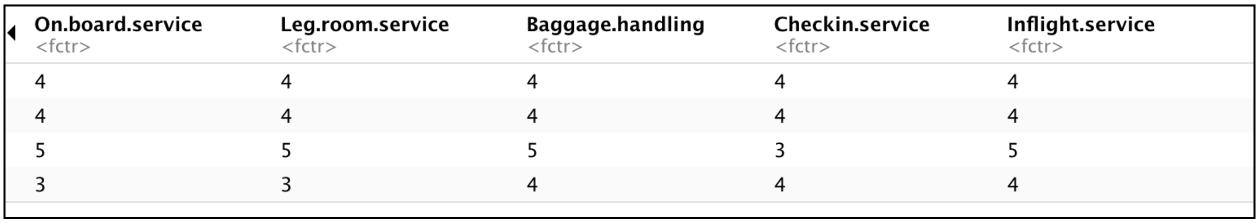
Gender <fctr>	Customer.Type <fctr>	Age <dbl>	Type.of.Travel <fctr>	Class <fctr>	Flight.Distance <dbl>
Female	Loyal Customer	42.56940	Business travel	Business	2768.9540
Male	Loyal Customer	38.43023	Business travel	Eco	616.5933
Male	Loyal Customer	40.03665	Business travel	Business	951.8293
Female	Loyal Customer	37.39852	Personal Travel	Eco	725.2364



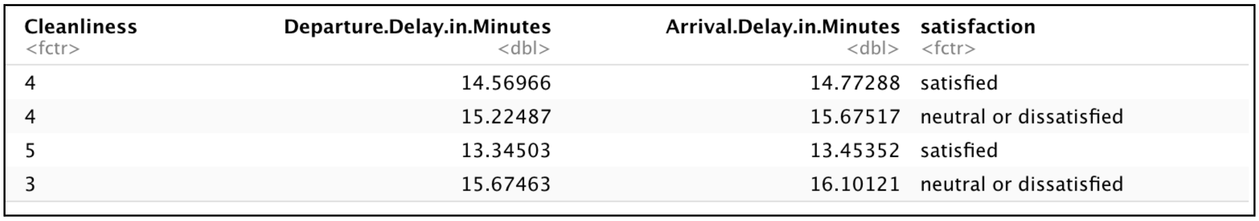
Inflight.wifi.service <fctr>	Departure.Arrival.time.convenient <fctr>	Ease.of.Online.booking <fctr>
2	5	2
2	2	2
1	1	1
3	4	3



Gate.location <fctr>	Food.and.drink <fctr>	Online.boarding <fctr>	Seat.comfort <fctr>	Inflight.entertainment <fctr>
3	4	4	4	4
2	4	4	4	4
1	5	5	5	5
3	3	3	3	3



On.board.service <fctr>	Leg.room.service <fctr>	Baggage.handling <fctr>	Checkin.service <fctr>	Inflight.service <fctr>
4	4	4	4	4
4	4	4	4	4
5	5	5	3	5
3	3	4	4	4



Cleanliness <fctr>	Departure.Delay.in.Minutes <dbl>	Arrival.Delay.in.Minutes <dbl>	satisfaction <fctr>
4	14.56966	14.77288	satisfied
4	15.22487	15.67517	neutral or dissatisfied
5	13.34503	13.45352	satisfied
3	15.67463	16.10121	neutral or dissatisfied