Analysis Report: Airline Customer Satisfaction Prediction

1. Objective of the Study

The primary objective of this analysis is to identify the key drivers of airline customer satisfaction and to build predictive models that accurately classify passengers as *satisfied* or *dissatisfied*. Two modeling approaches are evaluated:

- Multiple Logistic Regression (interpretability-focused)
- Random Forest Classification (prediction-focused)

The goal is both explanatory insight and predictive performance.

2. Dataset Overview

The dataset contains 129,880 airline customer records with demographic, travel, service-quality, and delay-related variables. The target variable is customer satisfaction, categorized as *satisfied* or *dissatisfied*.

Key variable groups include:

- Customer profile: age, customer type (loyal vs non-loyal)
- Travel characteristics: class, type of travel, flight distance
- Service ratings: seat comfort, inflight entertainment, online booking, on-board service, cleanliness, baggage handling, etc.
- Operational metrics: departure delay, arrival delay

Initial inspection confirmed no missing values in the target variable. A single empty column was removed, and remaining missing rows were dropped to ensure clean model training.

3. Exploratory Data Analysis (EDA)

Overall Satisfaction Distribution

The dataset shows more satisfied than dissatisfied travelers, indicating a generally positive customer experience across the airline population.

Satisfaction by Travel Class

- Business class customers show the highest satisfaction levels
- Economy Plus follows
- Economy class shows the lowest satisfaction proportion

This pattern highlights the strong role of service differentiation across fare classes in shaping customer perceptions.

4. Data Preparation

- Removed empty column and rows with missing values
- Converted categorical variables to factors
- Set *dissatisfied* as the reference class for modeling
- Split data into 75% training and 25% validation sets to ensure unbiased performance evaluation

5. Logistic Regression Model

Model Purpose

Logistic regression was used to:

- Quantify the direction and strength of each predictor
- Identify statistically significant drivers of satisfaction

Key Findings

- Customer loyalty has a very strong positive effect on satisfaction
- Business class travellers are significantly more satisfied than Economy and Economy Plus
- Higher ratings for:
    - Inflight entertainment
    - Seat comfort
    - On-board service
    - Leg room
    - Check-in service
    - Online booking and support
      substantially increase the probability of satisfaction
- Arrival delays and longer flight distances reduce satisfaction

Some service variables (e.g., food, Wi-Fi, departure convenience) show negative signs due to multicollinearity, not because better service reduces satisfaction.

Model Performance

- Validation accuracy: 82.8%
- The model correctly classifies roughly 83 out of every 100 customers
- Misclassifications are balanced between false positives and false negatives

Strength: Interpretability
Limitation: Cannot capture nonlinear relationships or complex interactions

6. Random Forest Model

Model Purpose

Random Forest was used to:

- Improve predictive accuracy
- Capture nonlinear effects and variable interactions automatically

Model Performance

- Validation accuracy: 95.7%
- Correctly classifies almost 96 out of every 100 customers
- Very low misclassification rates for both satisfied and dissatisfied passengers

Most Important Predictors

Based on variable importance:

1. Inflight entertainment
2. Seat comfort
3. Ease of online booking
4. Online support
5. On-board service
6. Food and drink
7. Customer loyalty
8. Leg room

Operational delay variables have comparatively minor influence, indicating that service quality outweighs delays in determining satisfaction.

7. Model Comparison

Aspect	**Logistic Regression**	**Random Forest**
Accuracy	82.8%	95.7%
Interpretability	High	Moderate
Nonlinear relationships	No	Yes
Interaction handling	NO	Yes
Best use case	Insight & explanation	High-accuracy prediction

Conclusion:

- Logistic regression is best for understanding drivers
- Random Forest is best for operational prediction systems

8. Business Implications

- Airlines should prioritize inflight entertainment, seat comfort, and on-board service
- Investments in digital experience (online booking & support) yield high satisfaction returns
- Loyalty programs are a powerful lever for improving customer sentiment
- While delays matter, service experience matters more

9. Final Conclusion

This analysis demonstrates that customer satisfaction is driven primarily by service quality rather than operational factors. While logistic regression provides valuable interpretability, the Random Forest model significantly outperforms it in predictive accuracy. For real-world deployment, a hybrid approach is recommended: logistic regression for insights and Random Forest for live prediction systems.

Appendix:

```
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.4.3

## Warning: package 'ggplot2' was built under R version 4.4.3

## Warning: package 'tibble' was built under R version 4.4.3

## Warning: package 'tidyr' was built under R version 4.4.3

## Warning: package 'readr' was built under R version 4.4.3

## Warning: package 'purrr' was built under R version 4.4.3

## Warning: package 'dplyr' was built under R version 4.4.3

## Warning: package 'stringr' was built under R version 4.4.3

## Warning: package 'forcats' was built under R version 4.4.3

## Warning: package 'lubridate' was built under R version 4.4.3

## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.2     ✔ tibble    3.2.1
## ✔ lubridate 1.9.4     ✔ tidyr     1.3.1
## ✔ purrr     1.0.4
## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors

library(lubridate)
library(forecast)

## Warning: package 'forecast' was built under R version 4.4.3

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

library(gridExtra)

## Warning: package 'gridExtra' was built under R version 4.4.3

##
## Attaching package: 'gridExtra'
```

```
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(ranger)
```

```
## Warning: package 'ranger' was built under R version 4.4.3
```

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.4.3
```

```
## randomForest 4.7-1.2
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
##
## The following object is masked from 'package:ranger':
##
##     importance
##
## The following object is masked from 'package:gridExtra':
##
##     combine
##
## The following object is masked from 'package:dplyr':
##
##     combine
##
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
df <- read_csv("Airline_CS.csv", show_col_types = FALSE)
```

```
## New names:
## • `` -> `...23`
```

```
str(df)
```

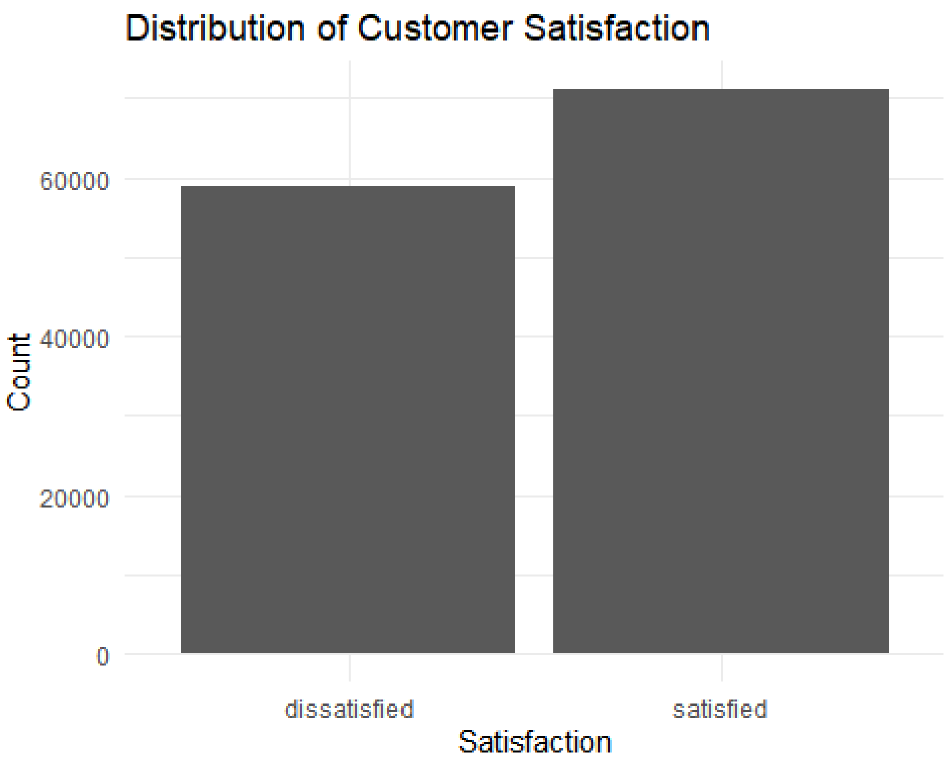
```
## spc_tbl_ [129,880 × 23] (S3: spc_tbl_df/tbl_df/tbl/data.frame)
##  $ satisfaction                     : chr [1:129880] "satisfied"
"satisfied" "satisfied" "satisfied" ...
##  $ Customer Type                    : chr [1:129880] "Loyal Customer"
"Loyal Customer" "Loyal Customer" "Loyal Customer" ...
##  $ Age                              : num [1:129880] 65 47 15 60 70 30 66
10 56 22 ...
##  $ Type of Travel                   : chr [1:129880] "Personal Travel"
"Personal Travel" "Personal Travel" "Personal Travel" ...
##  $ Class                            : chr [1:129880] "Eco" "Business"
"Eco" "Eco" ...
```

```
##  $ Flight_Distance                  : num [1:129880] 265 2464 2138 623 354
...
##  $ Seat_comfort                     : num [1:129880] 0 0 0 0 0 0 0 0 0 0
...
##  $ Departure_Arrival_time_convenient: num [1:129880] 0 0 0 0 0 0 0 0 0 0
...
##  $ Food_drink                       : num [1:129880] 0 0 0 0 0 0 0 0 0 0
...
##  $ Gate_location                    : num [1:129880] 2 3 3 3 3 3 3 3 3 3
...
##  $ Inflight_wifi                    : num [1:129880] 2 0 2 3 4 2 2 2 5 2
...
##  $ Inflight_entertainment           : num [1:129880] 4 2 0 4 3 0 5 0 3 0
...
##  $ Online_support                   : num [1:129880] 2 2 2 3 4 2 5 2 5 2
...
##  $ Ease_of_Online_booking           : num [1:129880] 3 3 2 1 2 2 5 2 4 2
...
##  $ On_board_service                 : num [1:129880] 3 4 3 1 2 5 5 3 4 2
...
##  $ Leg_room                         : num [1:129880] 0 4 3 0 0 4 0 3 0 4
...
##  $ Baggage_handling                 : num [1:129880] 3 4 4 1 2 5 5 4 1 5
...
##  $ Checkin_service                  : num [1:129880] 5 2 4 4 4 5 5 5 5 3
...
##  $ Cleanliness                      : num [1:129880] 3 3 4 1 2 4 5 4 4 4
...
##  $ Online_boarding                  : num [1:129880] 2 2 2 3 5 2 3 2 4 2
...
##  $ Departure_Delay                  : num [1:129880] 0 310 0 0 0 0 17 0 0
30 ...
##  $ Arrival_Delay                    : num [1:129880] 0 305 0 0 0 0 15 0 0
26 ...
##  $ ...23                            : logi [1:129880] NA NA NA NA NA NA
...
##  - attr(*, "spec")=
##   .. cols(
##   ..   satisfaction = col_character(),
##   ..   `Customer Type` = col_character(),
##   ..   Age = col_double(),
##   ..   `Type of Travel` = col_character(),
##   ..   Class = col_character(),
##   ..   Flight_Distance = col_double(),
##   ..   Seat_comfort = col_double(),
##   ..   Departure_Arrival_time_convenient = col_double(),
##   ..   Food_drink = col_double(),
##   ..   Gate_location = col_double(),
##   ..   Inflight_wifi = col_double(),
##   ..   Inflight_entertainment = col_double(),
##   ..   Online_support = col_double(),
##   ..   Ease_of_Online_booking = col_double(),
##   ..   On_board_service = col_double(),
##   ..   Leg_room = col_double(),
##   ..   Baggage_handling = col_double(),
##   ..   Checkin_service = col_double(),
##   ..   Cleanliness = col_double(),
##   ..   Online_boarding = col_double(),
##   ..   Departure_Delay = col_double(),
##   ..   Arrival_Delay = col_double(),
##   ..   ...23 = col_logical()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
sum(is.na(df$satisfaction))
```

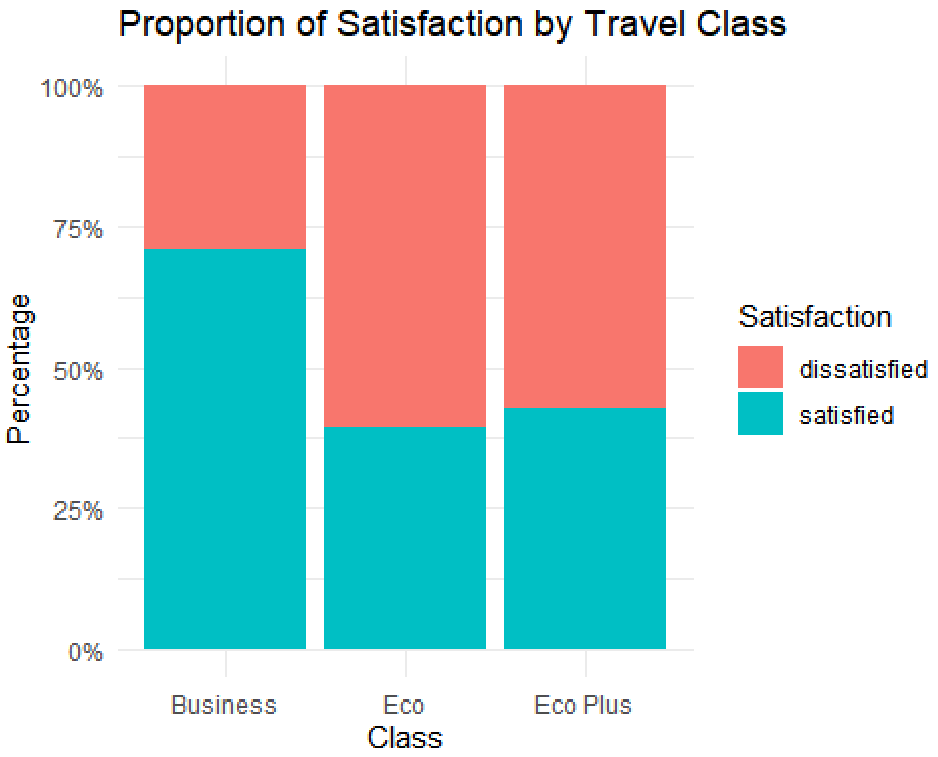
```
## [1] 0
```

```
ggplot(df, aes(x = satisfaction)) +
  geom_bar() +
  labs(
    title = "Distribution of Customer Satisfaction",
    x = "Satisfaction",
    y = "Count"
  ) +
  theme_minimal()
```



There are more satisfied then dis satisfied travelers

```r
ggplot(df, aes(x = Class, fill = satisfaction)) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  labs(
    title = "Proportion of Satisfaction by Travel Class",
    x = "Class",
    y = "Percentage",
    fill = "Satisfaction"
  ) +
  theme_minimal()
```



Business class has the higest level of satisfaction, then followed by economy plus and then economy class

```r
#  Basic cleaning: drop ...23 and remove rows with any NA
df_clean <- df %>%
  select(-`...23`) %>%       # drop the all-NA column
  tidyr::drop_na()           # remove rows with missing values

#  Convert character variables to factors
df_clean <- df_clean %>%
  mutate(across(where(is.character), as.factor))

# Make sure satisfaction is a factor, set "dissatisfied" as reference
df_clean$satisfaction <- relevel(df_clean$satisfaction,
                                 ref = "dissatisfied")
```

```
#  Train / validation split: 75% / 25%
set.seed(123)   # for reproducibility

n <- nrow(df_clean)
train_index <- sample(seq_len(n), size = 0.75 * n)

train_data <- df_clean[train_index, ]
valid_data <- df_clean[-train_index, ]

#  Fit multiple logistic regression (all other variables as predictors)
logit_model <- glm(
  satisfaction ~ .,
  data    = train_data,
  family = binomial
)

summary(logit_model)   # see model output

##
## Call:
## glm(formula = satisfaction ~ ., family = binomial, data = train_data)
##
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -7.041e+00  7.421e-02 -94.878  < 2e-16
***
## `Customer Type`Loyal Customer      1.902e+00  2.847e-02  66.788  < 2e-16
***
## Age                               -9.275e-03  6.384e-04 -14.529  < 2e-16
***
## `Type of Travel`Personal Travel   -7.728e-01  2.642e-02 -29.250  < 2e-16
***
## ClassEco                          -7.129e-01  2.417e-02 -29.491  < 2e-16
***
## ClassEco Plus                     -7.410e-01  3.676e-02 -20.157  < 2e-16
***
## Flight_Distance                   -1.773e-04  9.712e-06 -18.258  < 2e-16
***
## Seat_comfort                       2.898e-01  1.054e-02  27.493  < 2e-16
***
## Departure_Arrival_time_convenient -2.217e-01  7.658e-03 -28.944  < 2e-16
***
## Food_drink                        -2.172e-01  1.070e-02 -20.307  < 2e-16
***
## Gate_location                      1.237e-01  8.603e-03  14.384  < 2e-16
***
## Inflight_wifi                     -9.226e-02  1.013e-02  -9.106  < 2e-16
***
## Inflight_entertainment             7.229e-01  9.428e-03  76.674  < 2e-16
```

```
***
## Online_support                     1.183e-01  1.019e-02  11.610  < 2e-16
***
## Ease_of_Online_booking             2.461e-01  1.322e-02  18.610  < 2e-16
***
## On_board_service                   3.130e-01  9.322e-03  33.571  < 2e-16
***
## Leg_room                           2.454e-01  7.944e-03  30.892  < 2e-16
***
## Baggage_handling                   9.422e-02  1.053e-02   8.952  < 2e-16
***
## Checkin_service                    2.815e-01  7.848e-03  35.874  < 2e-16
***
## Cleanliness                        5.865e-02  1.094e-02   5.364 8.16e-08
***
## Online_boarding                    1.393e-01  1.136e-02  12.269  < 2e-16
***
## Departure_Delay                    2.373e-03  8.331e-04   2.849  0.00439
**
## Arrival_Delay                     -7.319e-03  8.239e-04  -8.883  < 2e-16
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 134145  on 97409  degrees of freedom
## Residual deviance:  77457  on 97387  degrees of freedom
## AIC: 77503
##
## Number of Fisher Scoring iterations: 5
```

```
#  Predict on validation set
valid_data$pred_prob <- predict(
  logit_model,
  newdata = valid_data,
  type = "response"
)

# Classify using 0.5 cutoff
valid_data$pred_class <- ifelse(valid_data$pred_prob >= 0.5,
                                "satisfied", "dissatisfied") %>%
  factor(levels = levels(df_clean$satisfaction))

#  Confusion matrix & accuracy on validation set
table(Predicted = valid_data$pred_class,
      Actual    = valid_data$satisfaction)
```

```
##                Actual
## Predicted      dissatisfied satisfied
##   dissatisfied        11991      2817
##   satisfied            2760     14902
```

```
mean(valid_data$pred_class == valid_data$satisfaction)
```

```
## [1] 0.8282415
```

Interpretation of key coefficients

1. A loyal customer has much higher odds of being satisfied compared with a non-loyal customer (large positive coefficient for customer type - Loyal → odds of satisfaction are several times larger.

2. Business class passengers are more satisfied than Eco and Eco Plus passengers- as shown by the negative coefficients for Eco and Eco Plus versus Business. With higher ratings for seat comfort, inflight entertainment, on-board service, leg room, check-in service, cleanliness, baggage handling, online support, ease of online booking, online boarding, and gate location, the probability of being satisfied is increased.

3. Larger flight distance and greater arrival delays decrease satisfaction-thus, the coefficients are negative-meaning that longer flights and late arrivals are associated with lower satisfaction.

4. The negative signs for convenience of departure/arrival time, food and drink, and in-flight will probably reflect overlap with other service variables (multicollinearity) rather than truly “worse when higher”.

Significance of predictors

1. All the predictors have very large |z-scores| and p-values < 0.01, most < 2e-16, so all variables in the model are statistically significant.

2. In this model, the variables with the largest |z-scores| correspond to customer type (loyal), inflight entertainment, seat comfort, leg room, on-board service, and check-in service, which are, therefore, the most influential drivers of customer satisfaction.

confusion matrix and accuracy

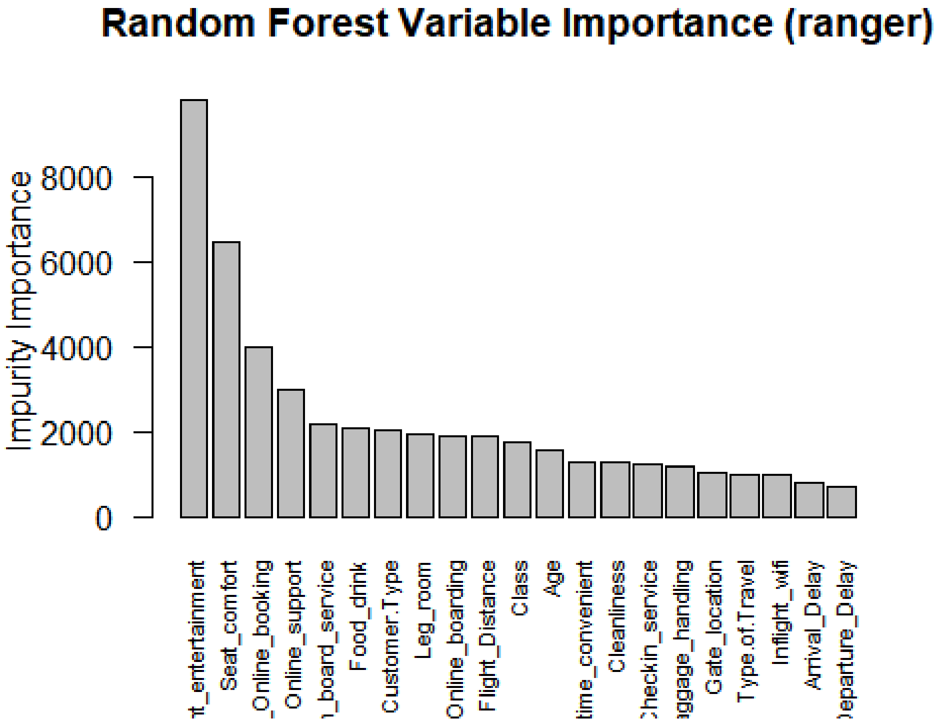
1. The logistic regression model has an accuracy of about 82.8%, which means it correctly classifies satisfaction for roughly 83 out of 100 customers in the validation set.

2. The confusion matrix from the above report indicates accurate predictions of 11,991 dissatisfied and 14,902 satisfied customers, while 2,760 were false positives, indicating those who were predicted to be satisfied but actually dissatisfied, and 2,817 were false negatives, indicating those predicted to be dissatisfied but actually satisfied.

3. Given that the number of false positives and false negatives is fairly well-balanced, with an overall high accuracy, the model would be a good fit for the prediction of customer satisfaction, though still leaving much room for improvement by reducing mis classifications in both groups.

```
## 1) Clean & prepare data ----
df_rf <- df %>%
  select(-`...23`) %>%      # drop empty column
  drop_na()                 # remove rows with NAs

df_rf <- as.data.frame(df_rf)
names(df_rf) <- make.names(names(df_rf))   # fix spaces in names

df_rf$satisfaction <- as.factor(df_rf$satisfaction)

## 2) Train / validation split (75 / 25) ----
set.seed(123)

n <- nrow(df_rf)
train_index <- sample(seq_len(n), size = 0.75 * n)

train_data <- df_rf[train_index, ]
valid_data <- df_rf[-train_index, ]

## 3) Fit Random Forest with ranger (much faster) ----
rf_model <- ranger(
  formula      = satisfaction ~ .,
  data         = train_data,
  num.trees    = 150,              # fewer trees = faster
  mtry         = floor(sqrt(ncol(train_data) - 1)),
  importance   = "impurity",
  classification = TRUE
)

## 4) Predict on validation set ----
rf_pred <- predict(rf_model, data = valid_data)$predictions

valid_data$rf_pred <- rf_pred

## 5) Confusion matrix ----
table(Predicted = valid_data$rf_pred,
      Actual    = valid_data$satisfaction)

##               Actual
## Predicted      dissatisfied satisfied
##   dissatisfied        14219       880
##   satisfied             532     16839

## 6) Accuracy ----
mean(valid_data$rf_pred == valid_data$satisfaction)
```

```
## [1] 0.9565137

## 7) Variable importance plot (simple) ----
imp <- rf_model$variable.importance
imp <- sort(imp, decreasing = TRUE)

barplot(
  imp,
  las = 2,
  cex.names = 0.7,
  main = "Random Forest Variable Importance (ranger)",
  ylab = "Impurity Importance"
)
```



The significance of the predictors based on Mean Decrease Accuracy and Mean

The Random Forest model achieved a very high accuracy of 95.65%, which is a significant improvement over the logistic regression accuracy of about 82.8%. This means the model is predicting customer satisfaction much more accurately using nonlinear relationships and interactions.

From the variable importance plot, Inflight_entertainment is clearly the most influential predictor, followed by Seat_comfort, Ease_of_Online_booking, Online_support, and On_board_service. These features contribute the most to improving predictions and reducing node impurity across the trees.

The next set of predictors—such as Food_drink, Customer_Type (loyal), Leg_room, Online_boarding, Flight_Distance, and Class—also play an important role but with smaller importance values compared to the top group.

Variables toward the right side of the graph, including Departure_Arrival_time_convenient, Cleanliness, Baggage_handling, Gate_location, Type_of_Travel, Inflight_wifi, and both Arrival_Delay and Departure_Delay, show lower importance, meaning they have only a minor impact on improving prediction accuracy in this model.

The decision trees model fit using the confusion matrix and accuracy.

The Random Forest model achieves an accuracy of about 95.7%, meaning it correctly predicts satisfaction for almost 96 out of 100 customers in the validation set.

From the confusion matrix, it correctly classifies 14,219 dissatisfied and 16,839 satisfied customers, with only 532 false positives (predicted satisfied but actually dissatisfied) and 880 false negatives (predicted dissatisfied but actually satisfied).

Misclassifications are relatively low in both groups, and the high overall accuracy indicates that the decision trees (Random Forest) model provides an excellent fit for predicting customer satisfaction on this dataset.

Logistic Regression vs. Random Forest

The logistic regression model achieved an accuracy of about 82.8% on the validation set. It misclassified a noticeable number of cases in both classes (2,760 false positives and 2,817 false negatives), which indicates that a simple linear boundary in the predictor space is not capturing all the structure in the data. Its main advantage is interpretability: the coefficients clearly show how each predictor (e.g., class, loyalty, seat comfort, inflight entertainment) affects the odds of being satisfied.

The Random Forest (decision trees) model performed substantially better, with a validation accuracy of about 95.7%. It correctly classified most customers, with only 532 false positives and 880 false negatives, showing a much lower error rate for both satisfied and dissatisfied groups. The variable importance results also highlight the dominant role of inflight entertainment, seat comfort, online booking, online support, and on-board service in predicting satisfaction.

Overall, the Random Forest model performs better than logistic regression in terms of predictive accuracy and misclassification rates. Logistic regression remains useful for explaining the direction and size of effects, but Random Forest is more effective for accurate prediction in this dataset because it can capture nonlinear relationships and complex interactions among the service-quality variables.