Explaining Car vs Public Transport Choice Using Random Forests

Vera Gak Anagrova

10 December, 2025

Introduction

Commuters living outside metropolitan centres face a trade-off between using public transport or travelling by car. Understanding the determinants of this choice is essential for designing efficient and sustainable mobility policies. This study analyses 2,458 commuting trips in the Oslo region to investigate which factors influence an individual's likelihood of choosing the car over public transport. The dataset includes travel times by public transport and car, walking and waiting times, transfer requirements, service frequency, distance, and socio-demographic characteristics. To model mode choice, we estimate a random forest classifier and apply global interpretation methods to identify the most influential predictors. We focus on the four most important variables and evaluate whether they have a positive, negative or nonlinear relationship with car usage.

Methodology

We prepared the dataset by converting European decimal commas (e.g., time_pt, time_car, time_ratio, age) to standard numeric format and transforming the response (mode) and all dummy variables into factors. Rows with missing values (17 cases, <1%) were dropped, leaving 2,458 observations. We did not explicitly handle potential outliers, as random forests are generally robust to extreme observations due to their threshold-based split rules, and further preprocessing was unlikely to meaningfully affect results. Finally, the data were split into training (70%) and test (30%) sets using stratified sampling to preserve the class balance.

After preprocessing, we estimated a random forest classifier to model transportation mode choice. A random forest combines the predictions of many decision trees, each trained on a bootstrap sample while considering only a random subset of variables at each split. Intuitively, the method acts like a group of independent "voters," where the aggregated vote yields more stable and accurate predictions than any single tree. We trained a forest of 500 trees, since adding more trees beyond a few hundred rarely improves performance substantially but increases computation time. Model performance was then evaluated on the test set using accuracy, balanced accuracy, sensitivity, and specificity to account for class imbalance.

Because random forests consist of hundreds of interacting decision rules, they are considered "black box" models: their internal logic cannot be easily summarized or interpreted directly. To extract meaningful behavioural insights, we therefore rely on global interpretation techniques. First, variable importance based on the mean decrease in Gini identifies which predictors contribute most to distinguishing between car and public transport trips. Second, partial dependence plots (PDPs) illustrate how the predicted probability of choosing the car changes with a single predictor while holding all others constant, thereby revealing the direction, magnitude, and potential nonlinearity of each relationship.

Results

Figure 1 (Appendix A) displays the confusion matrix from the test set. The model correctly identifies most car users (621 correct predictions), but only a minority of public transport users (24 correct predictions). Misclassifications are asymmetric: 85 public transport users are misclassified as car, while only 7 car users are incorrectly predicted as public transport. This reflects the strong class imbalance, where car users dominate the sample and are therefore easier for the model to learn.

The random forest (500 trees, mtry = 3) also reports an out-of-bag (OOB) error rate of approximately 13.5%. The OOB confusion matrix shows the same pattern: it classifies car users very accurately (class error ≈ 2%), but struggles with public transport users (class error ≈ 80%). This confirms that the imbalance in the data makes public transport users harder to predict.

The test accuracy of 0.88 exceeds the no-information rate (i.e., the accuracy obtained by always predicting the majority class "Car"), indicating that the model learns meaningful patterns rather than simply exploiting class imbalance. Because accuracy alone can exaggerate performance when classes are imbalanced, we also consider balanced accuracy, which averages the performance over both travel modes. Balanced accuracy remains high, indicating that the classifier performs well even when accounting for the difficulty of detecting the minority class. Sensitivity for car users is very high, meaning the model successfully identifies most drivers, whereas specificity is lower, showing that some public transport users are misclassified as car users. This outcome is expected, as the model has many more examples from which to learn car travel patterns than public transport patterns.

Table 1: **Performance Metrics for Random Forest**

	Train.Accuracy	Test.Accuracy	Kappa	Sensitivity	Specificity	Balanced.Accuracy
Accuracy	0.984	0.875	0.297	0.989	0.22	0.605

To determine which variables influence travel decisions the most, we inspect random forest variable importance scores based on the mean decrease in Gini impurity. The four most influential predictors are time_ratio (public transport time relative to car time), followed by time_car, time_pt, and age. These results align with transport economic intuition: commuters are highly responsive to time-related costs, both in absolute terms (how long each mode takes) and relative terms (how much slower one mode is than the other). Socio-demographic characteristics such as income or education and service attributes such as frequency or transfer requirements have some predictive value, but their importance is notably lower than that of travel times, highlighting that mode choice is largely driven by perceived travel time convenience.

The partial dependence plots provide insight into how these key predictors influence decision-making (Figure 2, Appendix A). The effect of time_ratio is steep and nonlinear: when public transport is only slightly slower than driving, the probability of choosing the car remains moderate, but once public transport becomes substantially slower, the probability of driving increases sharply. Longer car travel time (time_car) reduces the likelihood of choosing the car, suggesting that congestion or distance discourages driving. In contrast, longer public transport travel time (time_pt) increases the likelihood of choosing the car, indicating that slow public services deter usage. Finally, age shows a mild positive trend, with older individuals more likely to drive, plausibly due to greater car ownership, stronger preference for comfort, or habitual travel behaviours, while younger commuters are relatively more inclined to use public transport.

Conclusion

The results show that the relative competitiveness of travel time is the dominant driver of mode choice in the Oslo region. Commuters strongly favour the car when public transport is substantially slower, whereas congestion that increases car travel time reduces its appeal. Car use also rises slightly with age, likely due to greater vehicle access and stronger preferences for comfort. These findings suggest that reducing travel time disadvantages of public transport is essential for discouraging car dependence. Policy efforts should prioritise faster and more efficient services, such as express routes with limited stops, dedicated bus lanes with signal priority, higher service frequency, and improved access to stops. Such measures are likely to yield disproportionate benefits where public transport currently lags behind the car, making sustainable modes more competitive for daily commuters.

Nevertheless, the analysis is subject to some limitations. The model does not include monetary factors such as fuel prices, ticket costs, or parking fees, which are known to influence mode choice. In addition, class imbalance reduces the model's ability to correctly identify public transport users, and because random forests are predictive rather than causal, they reveal statistical associations rather than the behavioural effects of policy interventions. Future work could incorporate cost variables, apply causal inference methods, or use class weighting or oversampling techniques to improve sensitivity for public transport users. Since these users are underrepresented and harder to model, targeted improvements to public transport travel times may help both actual commuting behaviour and future predictive performance.

Appendix A

Figure 1: Confusion Matrix Heatmap

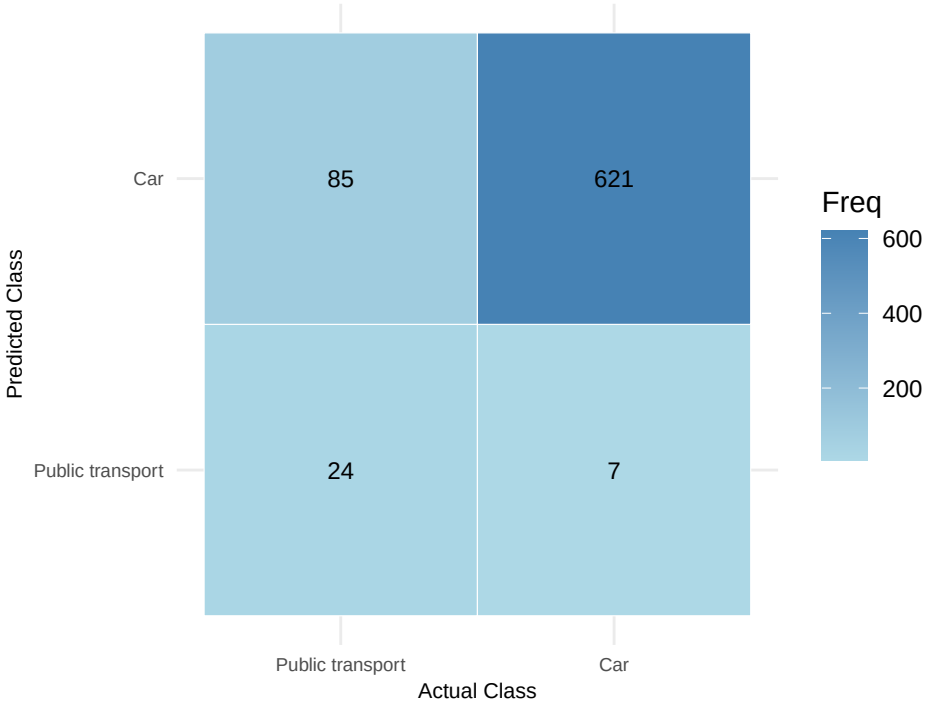
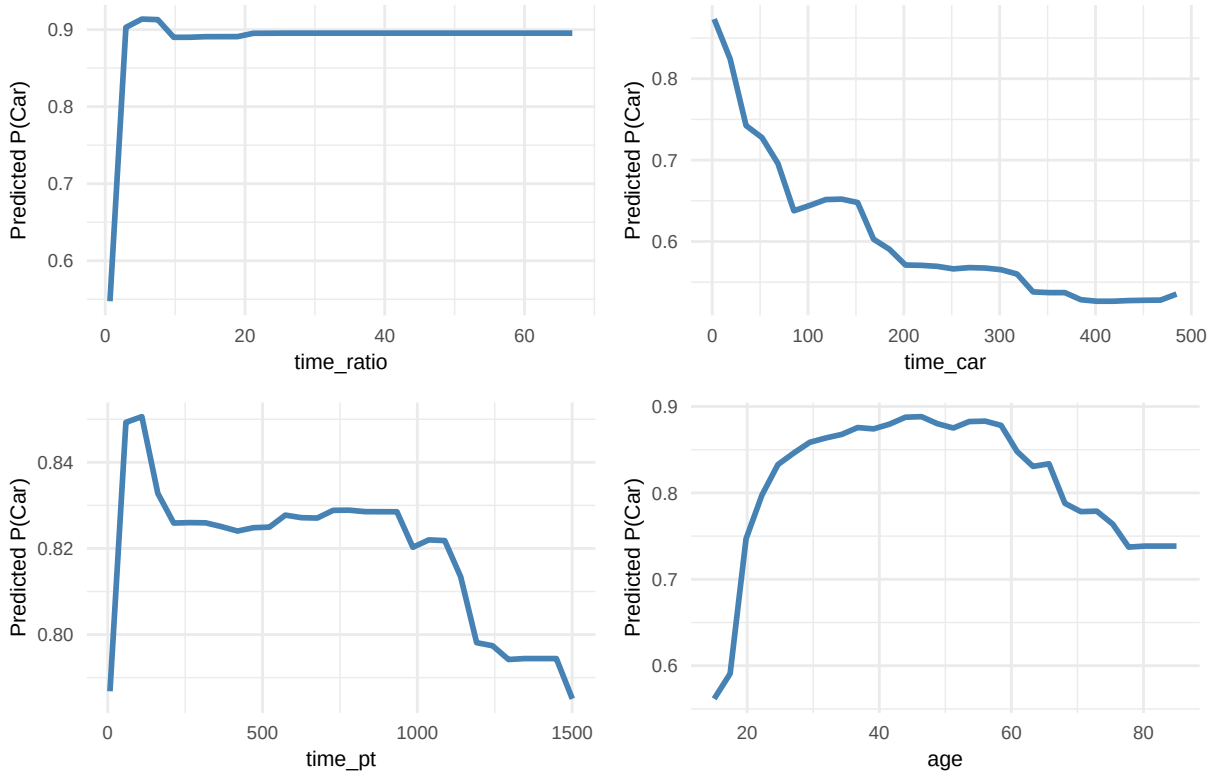


Figure 2: Partial Dependence of Top Predictors



Appendix B

```r
################################################################
# Load Libraries
################################################################
library(randomForest); library(caret); library(ggplot2)
library(patchwork); library(kableExtra); library(knitr)

################################################################
# Import and Clean Data
################################################################
load("TransportModeSweden.2526.RData")

# Inspection
str(data); head(data); summary(data)

# Convert numeric variables from comma to dot
num_cols <- c("time_pt", "time_car", "time_ratio", "age")

for (col in num_cols) {
  data[[col]] <- as.numeric(gsub(",", ".", data[[col]]))
}

# Convert outcome variable to factor
data$mode <- factor(
  data$mode,
  levels = c(0, 1),
  labels = c("Public transport", "Car"))

# Convert dummy variables to factor
dummy_vars <- c(
  "one_transfer", "mult_transfer", "walk_500", "wait_5",
  "high_freq", "dist_20", "high_inc", "high_ed", "woman")

data[dummy_vars] <- lapply(data[dummy_vars], factor)

# Check and remove missing values (17 rows)
colSums(is.na(data))
data <- na.omit(data)

# Final check
str(data)
summary(data)

################################################################
# Train/test split
################################################################
set.seed(123)  # for reproducibility

index <- createDataPartition(data$mode, p = 0.7, list = FALSE)
train <- data[index, ]
test  <- data[-index, ]

################################################################
# Train a Random Forest
################################################################
```

```r
set.seed(123)

rf_model <- randomForest(
  mode ~ .,
  data       = train,
  ntree      = 500,
  importance = TRUE)

rf_model
################################################################
# Predict on test set & performance
################################################################

# Confusion matrix (raw)
pred_test <- predict(rf_model, newdata = test)
cm <- confusionMatrix(pred_test, test$mode)

# Heatmap of confusion matrix
cm_data <- as.data.frame(cm$table)
colnames(cm_data) <- c("Predicted", "Actual", "Freq")

ggplot(cm_data, aes(Actual, Predicted, fill = Freq)) +
  geom_tile(color = "white") +
  geom_text(aes(label = Freq), size = 3) +
  scale_fill_gradient(low = "lightblue", high = "steelblue") +
  labs(
    title = "Figure 1: Confusion Matrix Heatmap",
    x = "Actual Class",
    y = "Predicted Class"
  ) +
  theme_minimal() +
  theme(axis.title.x = element_text(size = 8),axis.title.y = element_text(size = 8),
    axis.text.x  = element_text(size = 7),axis.text.y  = element_text(size = 7),
    plot.title   = element_text(size = 10, face = "bold"))

# Train & test confusion matrices (Car as positive class)
pred_train <- predict(rf_model, newdata = train)
cm_train   <- confusionMatrix(pred_train, train$mode, positive = "Car")

pred_test  <- predict(rf_model, newdata = test)
cm_test    <- confusionMatrix(pred_test, test$mode, positive = "Car")

# Overall performance metrics
stats <- data.frame(
  `Train Accuracy`    = cm_train$overall["Accuracy"],
  `Test Accuracy`     = cm_test$overall["Accuracy"],
  Kappa               = cm_test$overall["Kappa"],
  Sensitivity         = cm_test$byClass["Sensitivity"],
  Specificity         = cm_test$byClass["Specificity"],
  `Balanced Accuracy` = cm_test$byClass["Balanced Accuracy"]
)

kable(stats, digits = 3,
      caption = "Performance Metrics for Random Forest") %>%
  kable_styling(full_width = FALSE, position = "center")

################################################################
```

```r
# Global interpretation: variable importance
################################################################
varImpPlot(rf_model, type = 2,
           main = "Random forest variable importance")

imp        <- importance(rf_model, type = 2)  # MeanDecreaseGini
imp_sorted <- sort(imp[, "MeanDecreaseGini"], decreasing = TRUE)
imp_sorted

top4_names <- names(imp_sorted)[1:4]
top4_names

################################################################
# Global interpretation: partial dependence plots
# (effect of top 4 variables on P(choosing Car))
################################################################

get_pdp <- function(model, data, var, grid.size = 30) {
  stopifnot(var %in% names(data))

  x_vals <- seq(
    min(data[[var]], na.rm = TRUE),
    max(data[[var]], na.rm = TRUE),
    length.out = grid.size )

  tmp <- data
  pd  <- numeric(length(x_vals))

  for (i in seq_along(x_vals)) {
    tmp[[var]] <- x_vals[i]
    probs      <- predict(model, newdata = tmp, type = "prob")[, "Car"]
    pd[i]      <- mean(probs)
  }

  data.frame(x = x_vals, y = pd)
}

top_vars <- c("time_ratio", "time_car", "time_pt", "age")

pdp_list <- lapply(top_vars, function(v) {
  df_pdp <- get_pdp(rf_model, train, v)
  ggplot(df_pdp, aes(x = x, y = y)) +
    geom_line(color = "steelblue", size = 1) +
    labs(x = v, y = "Predicted P(Car)") +
    theme_minimal() +
    theme(axis.title.x = element_text(size = 8),axis.title.y = element_text(size = 8),
      axis.text.x  = element_text(size = 7),axis.text.y  = element_text(size = 7) )})

((pdp_list[[1]] | pdp_list[[2]]) /
 (pdp_list[[3]] | pdp_list[[4]])) +
  patchwork::plot_annotation(title = "Figure 2: Partial Dependence of Top Predictors") &
  theme(plot.title = element_text(size = 10, face = "bold"))
```

References

Altmeyer, P., C. C. S. Liem, and A. van Deursen. 2023. "Explaining Black-Box Models Through Counterfactuals." In *The Proceedings of the JuliaCon Conferences (JCON)*. https://resolver.tudelft.nl/uuid:446dc879-2782-4f89-9e25-120e912448ae. https://doi.org/10.21105/jcon.00130.

Boehmke, Bradley, and Brandon M. Greenwell. 2025. "Chapter 11: Random Forests." https://bradleyboehmke.github.io/HOML/random-forest.html.

GeeksforGeeks. 2025. "Random Forest Approach in r Programming." https://www.geeksforgeeks.org/r-language/random-forest-approach-in-r-programming/.

Investopedia. 2025. "What Is a Black Box Model? Definition, Uses, and Examples." https://www.investopedia.com/terms/b/blackbox.asp.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2023. *An Introduction to Statistical Learning: With Applications in r*. 2nd ed. New York: Springer.