Machine Learning Project

Dr.Öğr.Üy. MUSTAFA ÇAVUŞ

FEHMİCAN KORKUTER
Melih GÜNDÜZ

Problem, Features, and Target

Problem : Predicting hotel reservation cancellations using various features.

Features : number of adults, number of children, number of weekend,nights, type of meal plan, required car parking space, room type reserved, lead_time, arrival year, arrival month, arrival date, market segment type , repeated guest, number of previous cancellations, number of previous bookings not canceled, average price per room number of special requests, booking status

Target : The aim of the analysis is to create models according to the given features and predicting whether reservations are canceled or not and evaluating the performance of the models

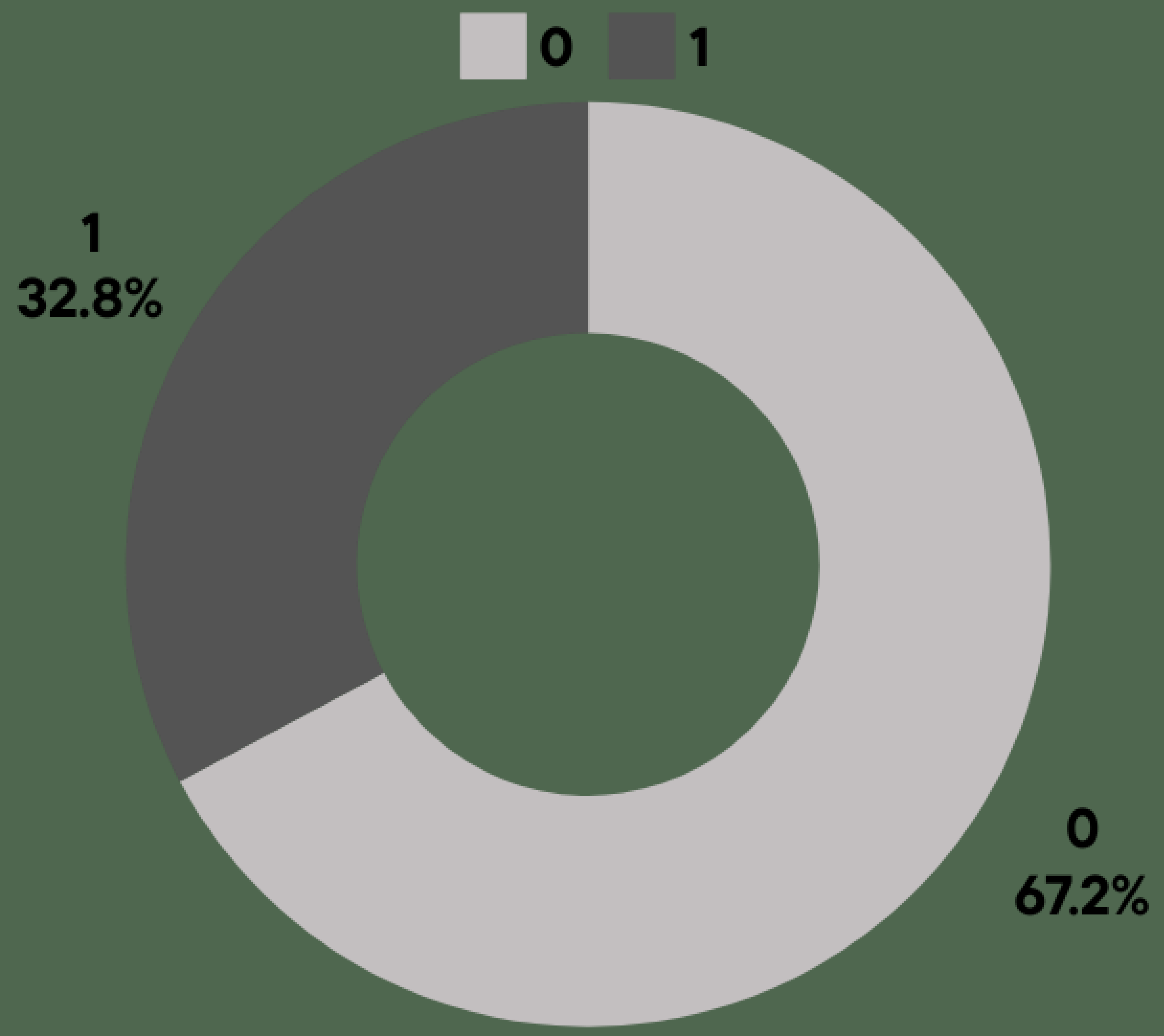
Dataset Overview

The dataset has a total of 36275 observations and 18 variables. Among these, 15 are numerical variables and 3 are categorical variables.

```
tibble [36,275 × 18] (S3: tbl_df/tbl/data.frame)
 $ no_of_adults                        : num [1:36275] 2 2 1 2 2 2 2 2 3 2 ...
 $ no_of_children                      : num [1:36275] 0 0 0 0 0 0 0 0 0 0 ...
 $ no_of_weekend_nights                : num [1:36275] 1 2 2 0 1 0 1 1 0 0 ...
 $ no_of_week_nights                   : num [1:36275] 2 3 1 2 1 2 3 3 4 5 ...
 $ type_of_meal_plan                   : chr [1:36275] "Meal Plan 1" "Not Selected"
"Meal Plan 1" "Meal Plan 1" ...
 $ required_car_parking_space          : num [1:36275] 0 0 0 0 0 0 0 0 0 0 ...
 $ room_type_reserved                  : chr [1:36275] "Room_Type 1" "Room_Type 1"
"Room_Type 1" "Room_Type 1" ...
 $ lead_time                           : num [1:36275] 224 5 1 211 48 346 34 83 121 44
...
 $ arrival_year                        : num [1:36275] 2017 2018 2018 2018 2018 ...
 $ arrival_month                       : num [1:36275] 10 11 2 5 4 9 10 12 7 10 ...
 $ arrival_date                        : num [1:36275] 2 6 28 20 11 13 15 26 6 18 ...
 $ market_segment_type                 : chr [1:36275] "Offline" "Online" "Online"
"Online" ...
 $ repeated_guest                      : num [1:36275] 0 0 0 0 0 0 0 0 0 0 ...
 $ no_of_previous_cancellations        : num [1:36275] 0 0 0 0 0 0 0 0 0 0 ...
 $ no_of_previous_bookings_not_canceled: num [1:36275] 0 0 0 0 0 0 0 0 0 0 ...
 $ avg_price_per_room                  : num [1:36275] 65 106.7 60 100 94.5 ...
 $ no_of_special_requests              : num [1:36275] 0 1 0 0 0 1 1 1 1 3 ...
 $ booking_status                      : num [1:36275] 0 0 1 1 1 1 0 0 0 0 ...
```

Check the imbalance problem

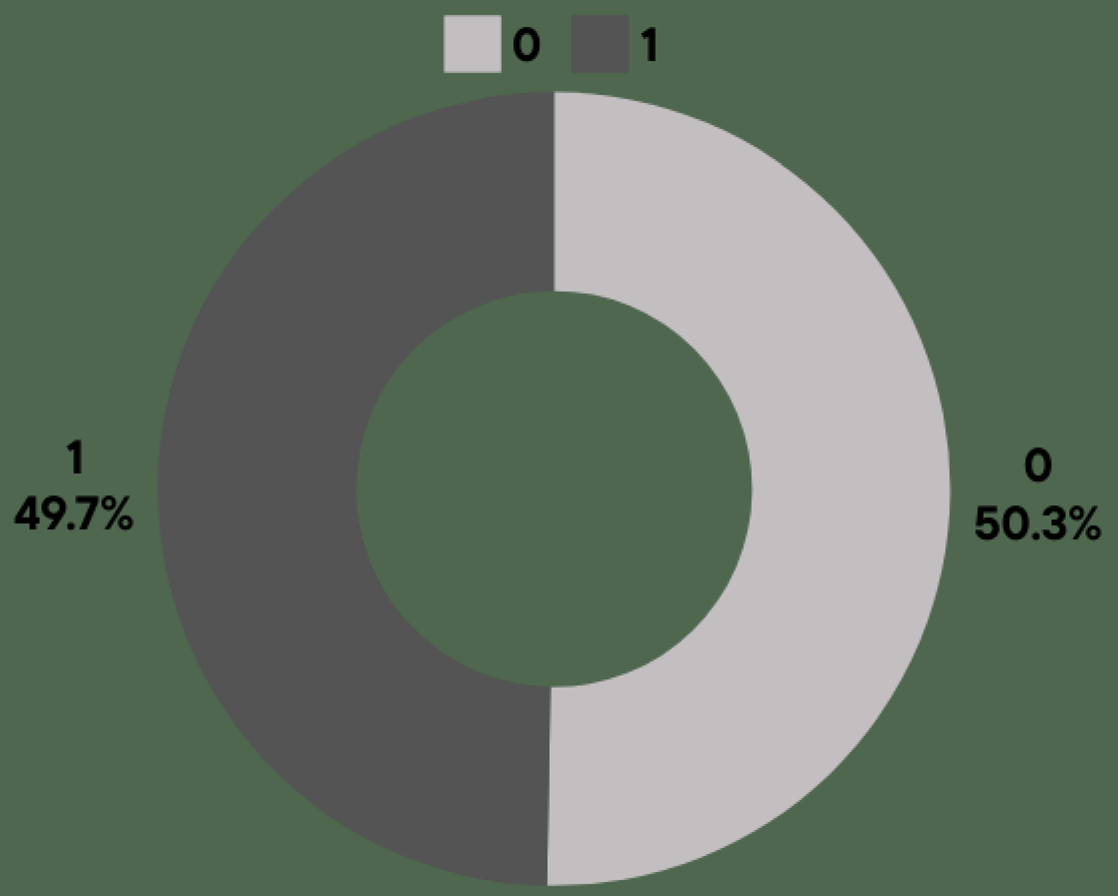
```
table(hotelnew$booking_status)/dim(hotelnew) [1]
```



Oversample

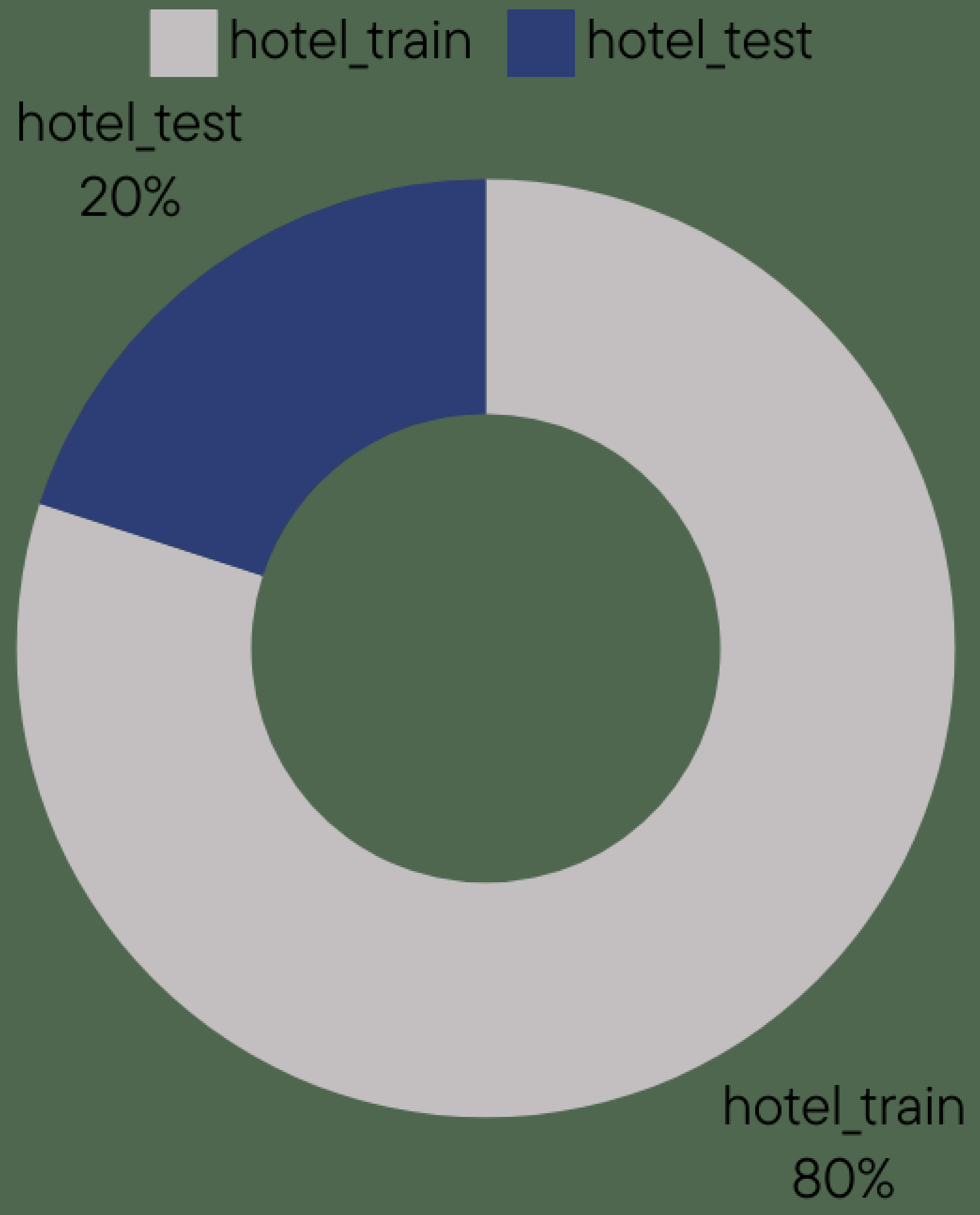
```
set.seed(123)
data_balanced_s <- ovun.sample(booking_status~., data = hotelnew,
                               method = "over", p=0.5)
data_balanced <- data_balanced_s$data

table(data_balanced$booking_status)/dim(data_balanced) [1]
```



Splitting The Dataset

```
hotel_split <- initial_split(data = data_balanced, prop = 0.80)
hotel_train <- hotel_split |> training()
hotel_test <- hotel_split |> testing()
```



Train a Logistic Regression Model

```r
lr_model <- glm(hotel_train$booking_status~., data=hotel_train,
                family = "binomial")
summary(lr_model)
```

```
Call:
glm(formula = hotel_train$booking_status ~ ., family = "binomial",
    data = hotel_train)

Coefficients:
                                       Estimate Std. Error z value Pr(>|z|)
(Intercept)                          -6.695e+02  9.018e+01  -7.425 1.13e-13 ***
no_of_children                        1.314e-01  4.656e-02   2.821  0.00479 **
no_of_weekend_nights                  1.503e-01  1.525e-02   9.855  < 2e-16 ***
no_of_week_nights                     5.260e-02  9.310e-03   5.650 1.61e-08 ***
type_of_meal_planMeal Plan 2          2.157e-01  5.220e-02   4.133 3.58e-05 ***
type_of_meal_planMeal Plan 3          1.150e+01  9.337e+01   0.123  0.90194
type_of_meal_planNot Selected         2.234e-01  4.121e-02   5.423 5.87e-08 ***
required_car_parking_space           -1.706e+00  1.029e-01 -16.579  < 2e-16 ***
room_type_reservedRoom_Type 2        -4.719e-01  1.027e-01  -4.593 4.38e-06 ***
room_type_reservedRoom_Type 3        -2.825e-01  1.073e+00  -0.263  0.79231
room_type_reservedRoom_Type 4        -2.048e-01  4.018e-02  -5.098 3.44e-07 ***
room_type_reservedRoom_Type 5        -6.642e-01  1.658e-01  -4.006 6.17e-05 ***
room_type_reservedRoom_Type 6        -9.016e-01  1.174e-01  -7.682 1.57e-14 ***
room_type_reservedRoom_Type 7        -1.223e+00  2.285e-01  -5.349 8.84e-08 ***
lead_time                             1.649e-02  2.152e-04  76.654  < 2e-16 ***
arrival_year                          3.309e-01  4.469e-02   7.406 1.31e-13 ***
arrival_month                        -4.651e-02  4.944e-03  -9.407  < 2e-16 ***
arrival_date                          6.620e-04  1.502e-03   0.441  0.65935
market_segment_typeComplementary     -1.901e+01  1.282e+02  -0.148  0.88212
market_segment_typeCorporate         -9.948e-01  2.060e-01  -4.830 1.37e-06 ***
market_segment_typeOffline           -2.002e+00  1.976e-01 -10.135  < 2e-16 ***
market_segment_typeOnline            -1.097e-01  1.950e-01  -0.563  0.57372
repeated_guest                       -2.298e+00  3.866e-01  -5.944 2.79e-09 ***
no_of_previous_cancellations          2.286e-01  5.254e-02   4.352 1.35e-05 ***
no_of_previous_bookings_not_canceled -1.046e-01  7.591e-02  -1.378  0.16820
avg_price_per_room                    1.699e-02  5.623e-04  30.221  < 2e-16 ***
no_of_special_requests               -1.474e+00  2.279e-02 -64.665  < 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: 53793  on 38803  degrees of freedom
Residual deviance: 35394  on 38777  degrees of freedom
AIC: 35448

Number of Fisher Scoring iterations: 15
```

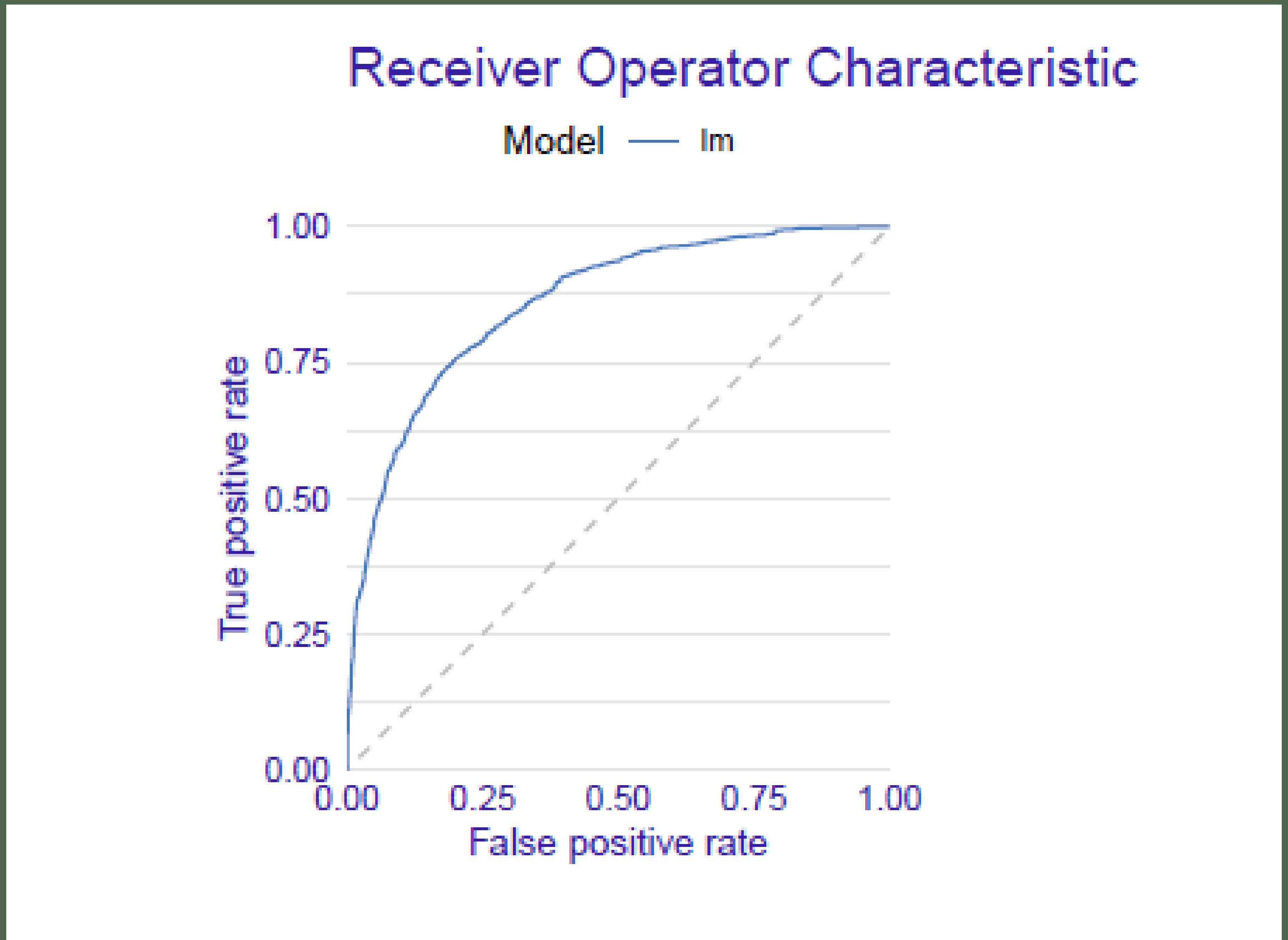
Confusion Matrix

```
confusionMatrix(table(ifelse(hotel_test$booking_status == "1", "1", "0"),
                       hotel_classes), positive= "1")
```

Confusion Matrix and Statistics		
hotel_classes		
	0	**1**
0	3825	1082
1	1097	3697

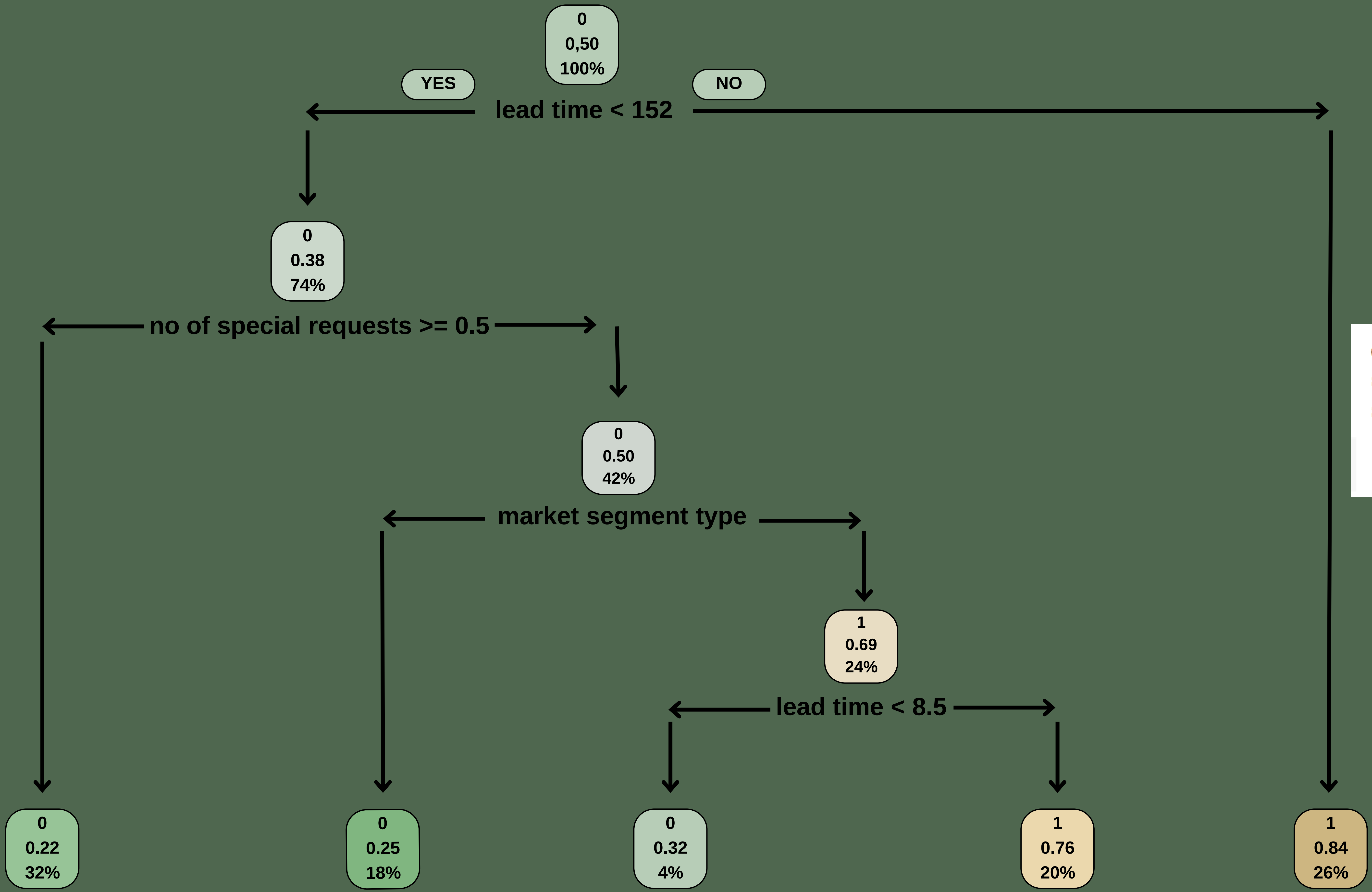
Accuracy	0.7754	**Pos Pred Value**	0.7712
95% CI	0.7969-0.7837	**Neg Pred Value**	0.7795
No Information Rate	0.5074	**Prevalence**	0.4926
P- Value [Acc > NIR]	<2e-16	**Detection Rate**	0.3811
Kappa	0.5507	**Detection Prevalence**	0.4942
Mcnemar's Test P-Value	0.7642	**Balanced Accuracy**	0.7754
Sensitivity	0.7736	**'Positive' Class**	1
Specificity	0.7771		

ROC Curve



Measures for	classification
recall	0.7711723
precision	0.7735928
f1	0.7723807
accuracy	0.775384
auc	0.8601501

Decision Tree




```
dt_model <- decision_tree() |>
set_engine("rpart") |>
set_mode("classification")

rpart.plot(dt_hotel$fit)
```

Confusion Matrix

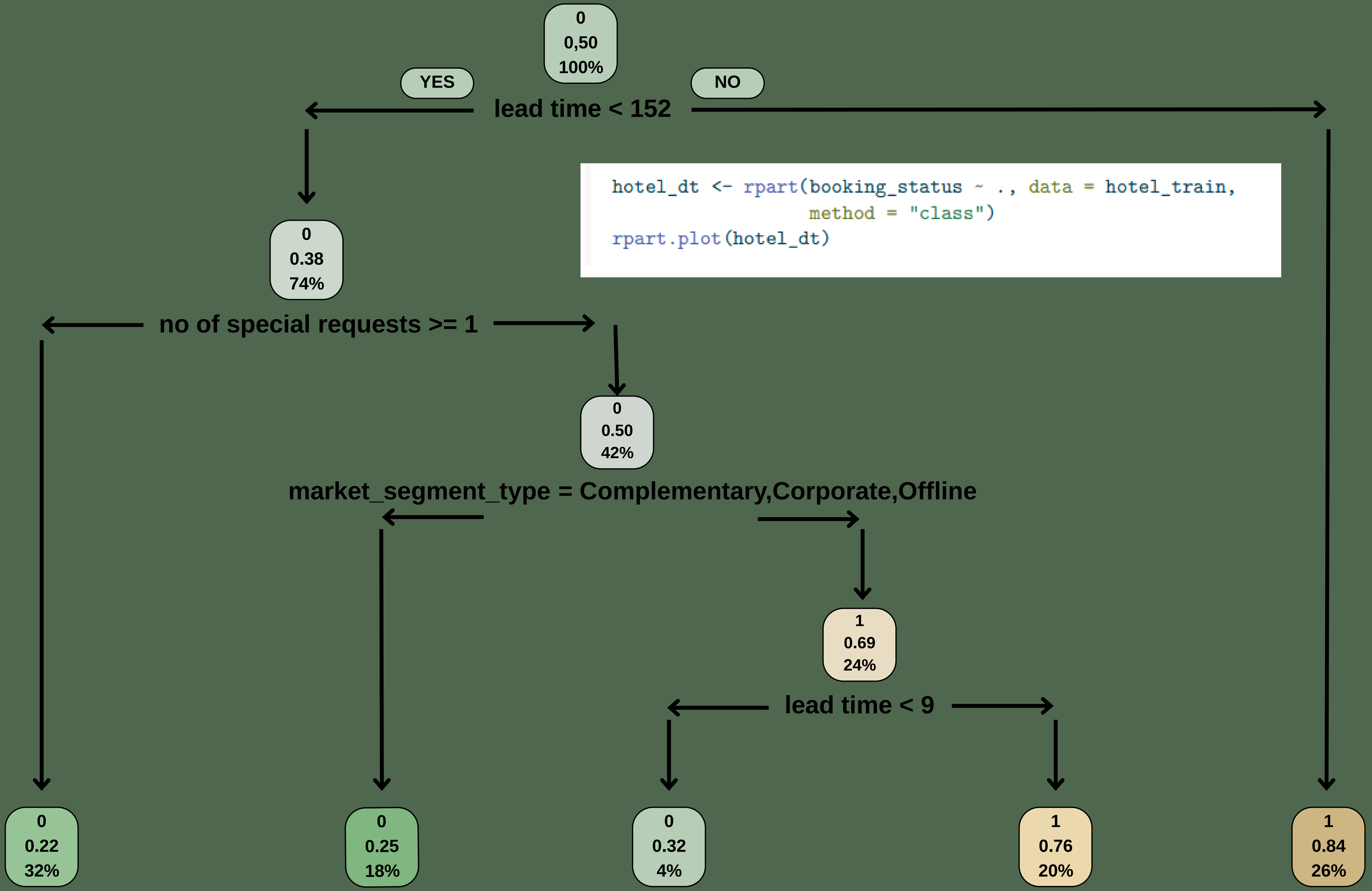
Truth		
Prediction	0	1
0	4029	1254
1	878	3540

Metric	estimator	estimate
accuracy	binary	0.780

Metric	estimator	estimate
sensivity	binary	0.821

Metric	estimator	estimate
f_meas	binary	0.791

The Overfitting Problem This code collects the necessary information to measure the performance of the decision tree model in the dataset.

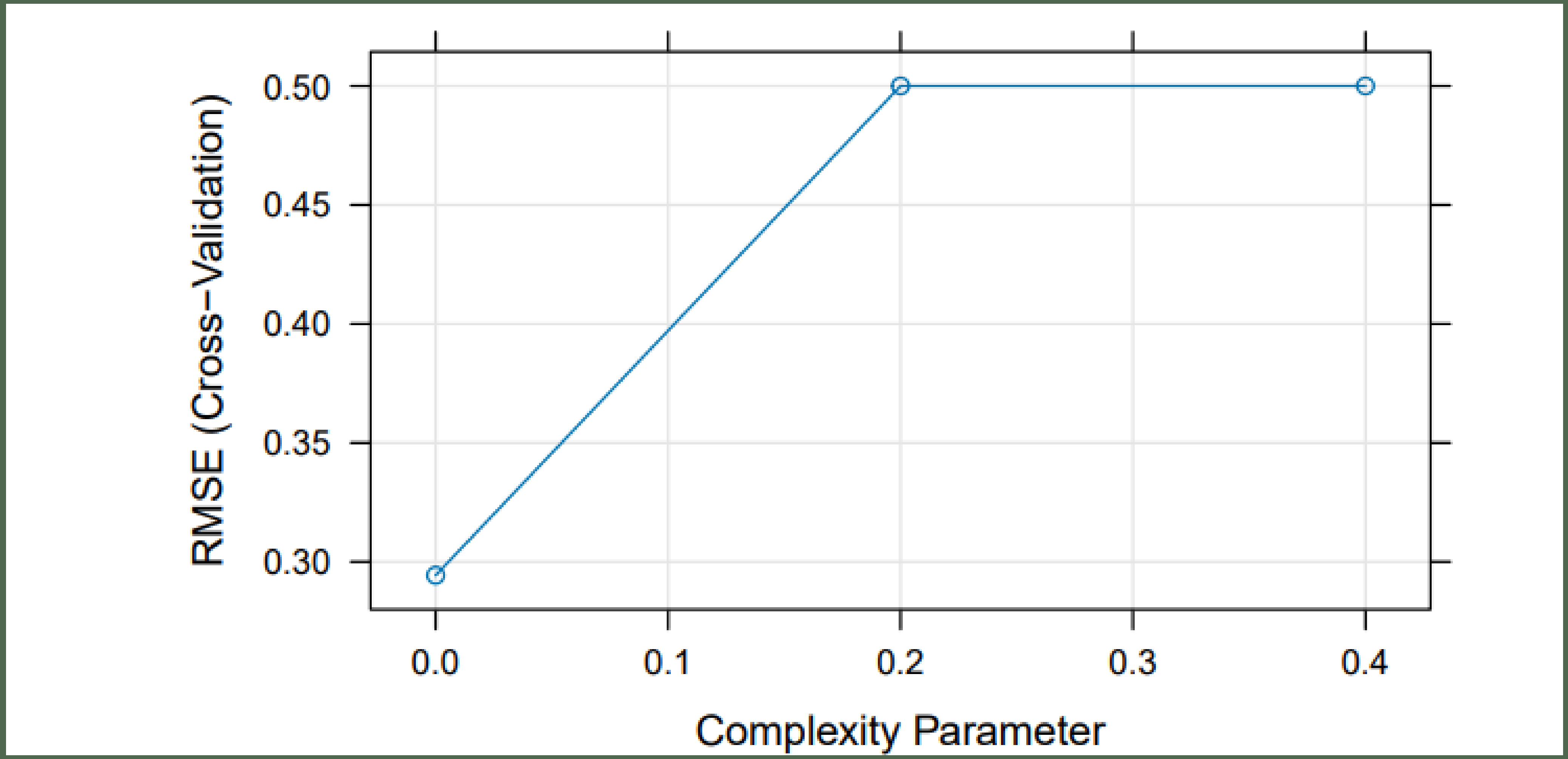



```r
hotel_dt <- rpart(booking_status ~ ., data = hotel_train,
                  method = "class")
rpart.plot(hotel_dt)
```

Improve The Prediction Performance Of The Decision Tree

Model Tuning Hyperparameters (Grid Search in Caret)

```r
fit_control <- trainControl(method = "cv", number = 10)
hyp_dt_model <- train(booking_status ~ .,
                      data = hotel_train,
                      method = "rpart",
                      trControl = fit_control,
                      tuneGrid = expand.grid(cp = seq(0, 0.5, 0.20)),
                      maxdepth =30,
                      cp = 0.01)
plot(hyp_dt_model)
```



Training Bagging Model

Type	Regression
Number of trees	500
Sample size	38804
Number of independent variables	17
Mtry	8
Target node size	5
Variable importance mode	none
Splitrule	variance
OOB prediction error (MSE):	0.04548852
R squared (OOB):	0.8180474

```r
bagging_model <- ranger(booking_status ~ .,
                        data = hotel_train,
                        mtry = 8)
bagging_model
```

Confusion Matrix

```
bagging_class_predict <- predict(bagging_model, hotel_test)$predictions
factor_rf <- (ifelse(bagging_class_predict > 0.5 ,1 ,0))
confusionMatrix(table(ifelse(hotel_test$booking_status == "1", "1", "0"),
                           factor_rf), positive= "1")
```

Confusion Matrix and Statistics		
factor_rf		
	0	**1**
0	4582	325
1	223	4571

Accuracy	0.9435	**Pos Pred Value**	0.9535
95% CI	0.9387, 0.948	**Neg Pred Value**	0.9338
No Information Rate	0.5047	**Prevalence**	0.5047
P- Value [Acc > NIR]	< 2e-16	**Detection Rate**	0.4712
Kappa	0.887	**Detection Prevalence**	0.4942
Mcnemar's Test P-Value	1.6e-05	**Balanced Accuracy**	0.9436
Sensitivity	0.9336	**'Positive' Class**	1
Specificity	0.9536		

Thank you