

Final Project

```
library(arrow)
```

```
##  
## Attaching package: 'arrow'  
  
## The following object is masked from 'package:utils':  
##  
##     timestamp
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --  
## v dplyr      1.1.3      v readr      2.1.4  
## v forcats    1.0.0      v stringr   1.5.0  
## v ggplot2    3.4.3      v tibble    3.2.1  
## v lubridate  1.9.3      v tidyr     1.3.0  
## v purrr      1.0.2  
  
## -- Conflicts ----- tidyverse_conflicts() --  
## x lubridate::duration() masks arrow::duration()  
## x dplyr::filter()       masks stats::filter()  
## x dplyr::lag()          masks stats::lag()  
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
```

```
# data_path <- "C:/Users/csg20/Downloads/"  
# applications <- read_feather(paste0(data_path, "app_data_starter_coded.feather"))  
  
data_path <- "/Users/aoluwoleotimi/Datasets/" # AO Changed file path in order to knit  
applications <- read_feather(paste0(data_path, "app_data_starter_coded.feather"))
```

clean the columns

```
columns_with_x_suffix <- grep("\\.x$", names(applications), value = TRUE)  
  
names(applications)[names(applications) %in% columns_with_x_suffix] <- sub("\\.x$", "", columns_with_x_suffix)  
  
applications <- applications %>% select(-ends_with(".y"))
```

Feature Engineering

```

#quarter
applications <- applications %>%
  mutate(
    quarter = paste0(year(filing_date), "/", quarter(filing_date)),
  )

# Aggregate applications by quarter and examiner
applications <- applications %>%
  group_by(quarter, examiner_id) %>%
  mutate(new_applications = n_distinct(application_number)) %>%
  ungroup()

applications <- applications %>%
  group_by(quarter, examiner_id) %>%
  mutate(ISSUED_applications = sum(disposal_type == "ISS" & !duplicated(application_number)))

applications <- applications %>%
  group_by(quarter, examiner_id) %>%
  mutate(abn_applications = sum(disposal_type == "ABN" & !duplicated(application_number)))

applications <- applications %>%
  group_by(quarter, examiner_id) %>%
  mutate(PEN_applications = sum(disposal_type == "PEND" & !duplicated(application_number)))

applications <- applications %>%
  group_by(quarter, examiner_art_unit) %>%
  mutate(examiner_art_unit_num = n_distinct(examiner_id)) %>%
  ungroup()

```

```

max_quarter <- "2017/1"

applications <- applications %>%
  filter(quarter <= max_quarter)

# separation_indicator
applications <- applications %>%
  group_by(examiner_id) %>%
  mutate(max_quarter_examiner = max(quarter)) %>%
  ungroup() %>%
  mutate(separation_indicator = if_else(max_quarter_examiner < max_quarter, 1, 0))

# to delete latest_date that is beyond the max_quarter
max_year <- "2018"

applications <- applications %>%
  mutate(latest_date = as.Date(latest_date),
    year_latest_date = year(latest_date)) %>%
  filter(year_latest_date <= max_year) %>%
  select(-year_latest_date)

```

```

#au_move_indicator
applications <- applications %>%
  group_by(examiner_id) %>%

```

```

mutate(au_move_indicator = if_else(examiner_art_unit != lag(examiner_art_unit), 1, 0)) %>%
ungroup()

applications <- applications %>%
  mutate(au_move_indicator = if_else(is.na(au_move_indicator), 0, au_move_indicator))

```

New Variable Creation Base on Presentation Feedback

```

applications <- applications %>%
  mutate(TC_1700 = if_else(tc == 1700, 1, 0),
         TC_1600 = if_else(tc == 1600, 1, 0),
         TC_2100 = if_else(tc == 2100, 1, 0),
         TC_2400 = if_else(tc == 2400, 1, 0))

applications_summary <- applications %>%
  group_by(examiner_id) %>%
  summarise(
    TC_1700_count = sum(TC_1700, na.rm = TRUE),
    TC_1600_count = sum(TC_1600, na.rm = TRUE),
    TC_2100_count = sum(TC_2100, na.rm = TRUE),
    TC_2400_count = sum(TC_2400, na.rm = TRUE),
    .groups = 'drop' # This drops the grouping structure, not required but cleaner
  )

# Step 2: Join this summary back to the original applications dataframe
applications <- applications %>%
  left_join(applications_summary, by = "examiner_id")

```

```

# get the count of au_moves by quarter
applications <- applications %>%
  group_by(examiner_id, quarter) %>%
  mutate(
    au_moves = sum(au_move_indicator)
  ) %>%
  ungroup()

#distinct uspc_class and uspc_subclass by quarter
applications <- applications %>%
  group_by(examiner_id, quarter) %>%
  mutate(
    num_classes = n_distinct(uspc_class),
    num_subclasses = n_distinct(uspc_subclass)
  ) %>%
  ungroup()

```

```

# get the process time
applications <- applications %>%
  mutate(
    patent_issue_date = as.Date(patent_issue_date),
    abandon_date = as.Date(abandon_date),
    processing_time = case_when(
      disposal_type == "ISS" ~ as.numeric(patent_issue_date - filing_date, units = "days"),

```

```

    disposal_type == "ABN" ~ as.numeric(abandon_date - filing_date, units = "days"),
    TRUE ~ 0
  )
)

applications <- applications %>%
  mutate(
    iss_time = ifelse(disposal_type == "ISS", processing_time, NA),
    abn_time = ifelse(disposal_type == "ABN", processing_time, NA)
  )

# Computing averages within the same dataframe
applications <- applications %>%
  group_by(examiner_id) %>%
  mutate(
    avg_processing = mean(processing_time, na.rm = TRUE),
    avg_ISS_processing = mean(iss_time, na.rm = TRUE),
    avg_ABN_processing = mean(abn_time, na.rm = TRUE)
  ) %>%
  ungroup()

```

Covariates Cleaning

```

columns_to_exclude <- c(
  "examiner_art_unit", "examiner_art_unit_num",
  "women_in_art_unit"
) # Due to a high number of missing vlaue in the gender value, the quality of the women_in_art_unit is

df <- applications[, !(names(applications) %in% columns_to_exclude)]
colSums(is.na(df))

```

```

## application_number      filing_date  examiner_name_last
##                0                0                0
## examiner_name_first examiner_name_middle  examiner_id
##                0                470360                0
##      uspc_class      uspc_subclass      patent_number
##                0                0                914727
## patent_issue_date      abandon_date      disposal_type
##      914257      1401229                0
## appl_status_code      appl_status_date      tc
##      4347      4348                0
##      gender      race      earliest_date
##      291954                0                0
##      latest_date      tenure_days      quarter
##                0                0                0
## new_applications ISSUED_applications  abn_applications
##                0                0                0
## PEN_applications max_quarter_examiner separation_indicator
##                0                0                0
## au_move_indicator      TC_1700      TC_1600
##                0                0                0

```

```
##          TC_2100          TC_2400          TC_1700_count
##          0          0          0
##      TC_1600_count      TC_2100_count      TC_2400_count
##          0          0          0
##          au_moves          num_classes          num_subclasses
##          0          0          0
##      processing_time          iss_time          abn_time
##          7          914257          1401238
##      avg_processing      avg_ISS_processing      avg_ABN_processing
##          0          12842          3720
```

```
#drop the na examiner_id rows
```

```
df <- subset(df, !is.na(examiner_id))
```

```
#aggregate to quarter
```

```
quarter_df <- df %>%
```

```
  group_by(examiner_id) %>%
```

```
  distinct(quarter, .keep_all = TRUE) %>%
```

```
  select(examiner_id, quarter, latest_date, separation_indicator, ISSUED_applications, PEN_applications
```

```
  arrange(examiner_id, quarter)
```

```
#collapse to individual observation
```

```
collapsed_df <- quarter_df %>%
```

```
  group_by(examiner_id) %>%
```

```
  summarize(
```

```
    gender = first(gender),
```

```
    race = first(race),
```

```
    tenure_days = first(tenure_days),
```

```
    ISSUED_applications = sum(ISSUED_applications),
```

```
    abandoned_applications = sum(abn_applications),
```

```
    au_moves = sum(au_moves),
```

```
    PEN_applications = sum(PEN_applications) / n(),
```

```
    separation_indicator = max(separation_indicator),
```

```
    num_classes = sum(num_classes),
```

```
    num_subclasses = sum(num_subclasses),
```

```
    avg_processing = first(avg_processing),
```

```
    avg_ISS_processing = first(avg_ISS_processing),
```

```
    avg_ABN_processing = first(avg_ABN_processing),
```

```
    TC_1700_count = first(TC_1700_count),
```

```
    TC_1600_count = first(TC_1600_count),
```

```
    TC_2100_count = first(TC_2100_count),
```

```
    TC_2400_count = first(TC_2400_count)
```

```
)
```

```
#append NA with 'unknown'
```

```
collapsed_df <- collapsed_df %>%
```

```
  mutate(gender = ifelse(is.na(gender), "unknown", gender))
```

```
#Save the collapsed data to csv
```

```
#write.csv(collapsed_df, "C:\\Users\\csg20\\OneDrive\\Desktop\\collapsed.csv", row.names=FALSE)
```

```
colSums(is.na(collapsed_df))
```

```
##          examiner_id          gender          race
##              0              0              0
## tenure_days ISSUED_applications abandoned_applications
##              0              0              0
##          au_moves          PEN_applications separation_indicator
##              0              0              0
## num_classes num_subclasses          avg_processing
##              0              0              0
## avg_ISS_processing avg_ABN_processing TC_1700_count
##          367          377              0
## TC_1600_count TC_2100_count TC_2400_count
##              0              0              0
```

```
#Fill NA to 0
```

```
collapsed_df <- collapsed_df %>%
  mutate_all(~ifelse(is.na(.), 0, .))
```

```
summary(collapsed_df)
```

```
## examiner_id          gender          race          tenure_days
## Min. :59012 Length:5627 Length:5627 Min. : 27
## 1st Qu.:66569 Class :character Class :character 1st Qu.:3098
## Median :75358 Mode :character Mode :character Median :4912
## Mean :78764 Mean :4430
## 3rd Qu.:93752 3rd Qu.:6091
## Max. :99990 Max. :6518
## ISSUED_applications abandoned_applications au_moves PEN_applications
## Min. : 0.0 Min. : 0.0 Min. : 0.00 Min. : 0.000
## 1st Qu.: 18.0 1st Qu.: 17.0 1st Qu.: 0.00 1st Qu.: 0.000
## Median : 105.0 Median : 67.0 Median : 0.00 Median : 1.167
## Mean : 191.8 Mean :105.2 Mean : 41.27 Mean : 1.578
## 3rd Qu.: 262.0 3rd Qu.:142.0 3rd Qu.: 59.00 3rd Qu.: 2.200
## Max. :1840.0 Max. :918.0 Max. :500.00 Max. :257.714
## separation_indicator num_classes num_subclasses avg_processing
## Min. :0.0000 Min. : 1.00 Min. : 1.0 Min. : -27.19
## 1st Qu.:0.0000 1st Qu.: 16.00 1st Qu.: 55.0 1st Qu.: 752.52
## Median :1.0000 Median : 47.00 Median : 193.0 Median : 996.16
## Mean :0.6707 Mean : 61.69 Mean : 260.1 Mean : 969.94
## 3rd Qu.:1.0000 3rd Qu.: 82.00 3rd Qu.: 375.0 3rd Qu.:1200.72
## Max. :1.0000 Max. :596.00 Max. :1804.0 Max. :3322.00
## avg_ISS_processing avg_ABN_processing TC_1700_count TC_1600_count
## Min. : 0.0 Min. : -690.2 Min. : 0.0 Min. : 0.00
## 1st Qu.: 983.5 1st Qu.: 885.5 1st Qu.: 0.0 1st Qu.: 0.00
## Median :1213.9 Median :1100.8 Median : 0.0 Median : 0.00
## Mean :1177.1 Mean :1053.3 Mean : 126.8 Mean : 92.49
## 3rd Qu.:1450.0 3rd Qu.:1306.1 3rd Qu.: 3.0 3rd Qu.: 0.00
## Max. :3078.0 Max. :3322.0 Max. :2372.0 Max. :2242.00
## TC_2100_count TC_2400_count
## Min. : 0.00 Min. : 0.00
## 1st Qu.: 0.00 1st Qu.: 0.00
```

```
## Median : 0.00 Median : 0.00
## Mean : 72.81 Mean : 62.12
## 3rd Qu.: 28.00 3rd Qu.: 1.00
## Max. :4228.00 Max. :5468.00
```

Model Building

```
collapsed_df$separation_indicator <- factor(make.names(as.character(collapsed_df$separation_indicator)))

collapsed_df$gender <- as.factor(collapsed_df$gender)
collapsed_df$race <- as.factor(collapsed_df$race)

#Train Split
library(caret)
```

```
## Loading required package: lattice
```

```
##
```

```
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
```

```
##
```

```
## lift
```

```
set.seed(123)
splitIndex <- createDataPartition(collapsed_df$separation_indicator, p = .75, list = FALSE)
train_data <- collapsed_df[splitIndex,]
test_data <- collapsed_df[-splitIndex,]

indexes <- sample(1:nrow(collapsed_df), size = 0.75 * nrow(collapsed_df))
train_data_lm <- collapsed_df[indexes, ]
test_data_lm <- collapsed_df[-indexes, ]
```

- What are the organizational and social factors associated with the length of patent application prosecution?

Hypothesis - The impact of tenure on outcomes like ISSUED applications, abandoned applications, or average processing times might differ between genders/races

```
train_control <- trainControl(method = "cv", number = 10)

#avg_processing as outcome
model <- train(avg_processing ~ tenure_days + num_classes + num_subclasses +
  ISSUED_applications + au_moves + gender + race +
  gender * tenure_days + TC_1700_count + TC_1600_count + TC_2100_count
  data = train_data_lm,
  method = "lm",
  trControl = train_control)
```

```
## Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
## attr(*, "non-estim") has doubtful cases
```

```

#avg_ISS_processing as outcome
model2 <- train(avg_ISS_processing ~ tenure_days + num_classes + num_subclasses+
  ISSUED_applications + au_moves + gender + race +
  gender * tenure_days + TC_1700_count + TC_1600_count +
  data = train_data_lm,
  method = "lm",
  trControl = train_control)

#avg_ABN_processing as outcome
model3 <- train(avg_ABN_processing ~ tenure_days + num_classes + num_subclasses+
  ISSUED_applications + au_moves + gender + race +
  gender * tenure_days + race * tenure_days,
  data = train_data_lm,
  method = "lm",
  trControl = train_control)

```

TC_21

TC_17

```
summary(model)
```

```

##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1054.76  -188.62    3.88   180.24  2679.22
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.256e+02  3.364e+01  12.649  < 2e-16 ***
## tenure_days       1.323e-01  7.199e-03  18.383  < 2e-16 ***
## num_classes       6.192e-01  1.569e-01   3.947  8.05e-05 ***
## num_subclasses    9.582e-02  6.270e-02   1.528  0.126553
## ISSUED_applications 3.412e-02  4.765e-02   0.716  0.473976
## au_moves          1.164e+00  1.048e-01  11.111  < 2e-16 ***
## gendermale        1.849e+01  3.054e+01   0.605  0.544999
## genderunknown     1.341e+01  4.199e+01   0.319  0.749379
## raceblack         1.727e+02  6.516e+01   2.651  0.008061 **
## raceHispanic      5.661e+01  6.241e+01   0.907  0.364424
## raceother         3.336e+03  2.034e+03   1.640  0.101184
## racewhite         1.102e+02  2.867e+01   3.844  0.000123 ***
## TC_1700_count     -6.317e-01  3.911e-02 -16.151  < 2e-16 ***
## TC_1600_count     -5.682e-01  3.936e-02 -14.437  < 2e-16 ***
## TC_2100_count     -3.564e-01  4.186e-02  -8.513  < 2e-16 ***
## TC_2400_count     -2.748e-01  3.678e-02  -7.472  9.55e-14 ***
## 'tenure_days:gendermale'  2.321e-03  6.291e-03   0.369  0.712177
## 'tenure_days:genderunknown' 1.694e-03  8.846e-03   0.192  0.848135
## 'tenure_days:raceblack'   -3.791e-02  1.377e-02  -2.754  0.005914 **
## 'tenure_days:raceHispanic' -2.013e-02  1.388e-02  -1.451  0.146982
## 'tenure_days:raceother'   -5.305e-01  3.469e-01  -1.529  0.126304
## 'tenure_days:racewhite'   -2.559e-02  6.067e-03  -4.217  2.52e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##

```



```
## Residual standard error: 306.3 on 4198 degrees of freedom
## Multiple R-squared:  0.3216, Adjusted R-squared:  0.3182
## F-statistic: 94.77 on 21 and 4198 DF,  p-value: < 2.2e-16
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2215.35  -195.19   35.03   232.76  1946.15
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.398e+02  4.087e+01  13.209  < 2e-16 ***
## tenure_days     1.408e-01  8.744e-03  16.099  < 2e-16 ***
## num_classes     1.220e+00  1.906e-01   6.402  1.70e-10 ***
## num_subclasses   7.489e-01  7.616e-02   9.833  < 2e-16 ***
## ISSUED_applications -1.244e+00  5.788e-02 -21.500  < 2e-16 ***
## au_moves         4.707e-01  1.273e-01   3.699  0.000219 ***
## gendermale       -3.884e+01  3.710e+01  -1.047  0.295237
## genderunknown     1.678e+00  5.100e+01   0.033  0.973751
## raceblack        -3.456e+01  7.915e+01  -0.437  0.662438
## raceHispanic      9.524e+01  7.581e+01   1.256  0.209061
## raceother         4.426e+03  2.471e+03   1.791  0.073352 .
## racewhite         5.233e+01  3.483e+01   1.503  0.133010
## TC_1700_count     -2.427e-01  4.750e-02  -5.108  3.40e-07 ***
## TC_1600_count     -3.524e-01  4.781e-02  -7.372  2.02e-13 ***
## TC_2100_count      1.109e-01  5.084e-02   2.181  0.029274 *
## TC_2400_count      9.248e-02  4.467e-02   2.070  0.038497 *
## 'tenure_days:gendermale' 1.511e-02  7.641e-03   1.977  0.048077 *
## 'tenure_days:genderunknown' 8.873e-03  1.074e-02   0.826  0.408931
## 'tenure_days:raceblack' 3.323e-04  1.672e-02   0.020  0.984146
## 'tenure_days:raceHispanic' -2.307e-02  1.686e-02  -1.369  0.171193
## 'tenure_days:raceother' -7.126e-01  4.214e-01  -1.691  0.090898 .
## 'tenure_days:racewhite' -1.650e-02  7.370e-03  -2.238  0.025249 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 372.1 on 4198 degrees of freedom
## Multiple R-squared:  0.3523, Adjusted R-squared:  0.3491
## F-statistic: 108.7 on 21 and 4198 DF,  p-value: < 2.2e-16
```

```
summary(model3)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -2087.3 -147.0 37.9 207.5 2676.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.596e+02  3.697e+01  15.136 < 2e-16 ***
## tenure_days     9.387e-02  7.911e-03  11.865 < 2e-16 ***
## num_classes     1.310e+00  1.724e-01   7.597 3.72e-14 ***
## num_subclasses   7.970e-01  6.890e-02  11.567 < 2e-16 ***
## ISSUED_applications -8.612e-01  5.236e-02 -16.446 < 2e-16 ***
## au_moves         5.704e-01  1.151e-01   4.953 7.58e-07 ***
## gendermale       3.610e+01  3.356e+01   1.075  0.2823
## genderunknown    2.852e+01  4.614e+01   0.618  0.5365
## raceblack        2.752e+01  7.161e+01   0.384  0.7007
## raceHispanic     1.484e+02  6.859e+01   2.164  0.0305 *
## raceother        2.156e+03  2.236e+03   0.965  0.3348
## racewhite        9.437e-02  3.151e+01   0.003  0.9976
## TC_1700_count    -3.436e-01  4.298e-02  -7.994 1.67e-15 ***
## TC_1600_count    -4.712e-01  4.325e-02 -10.893 < 2e-16 ***
## TC_2100_count     9.953e-02  4.600e-02   2.164  0.0306 *
## TC_2400_count     1.899e-01  4.042e-02   4.699 2.70e-06 ***
## 'tenure_days:gendermale'  6.100e-04  6.913e-03   0.088  0.9297
## 'tenure_days:genderunknown' 2.216e-03  9.721e-03   0.228  0.8197
## 'tenure_days:raceblack'    -1.353e-02  1.513e-02  -0.894  0.3713
## 'tenure_days:raceHispanic' -3.573e-02  1.525e-02  -2.343  0.0192 *
## 'tenure_days:raceother'    -3.444e-01  3.812e-01  -0.903  0.3663
## 'tenure_days:racewhite'    -9.842e-03  6.668e-03  -1.476  0.1400
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 336.6 on 4198 degrees of freedom
## Multiple R-squared:  0.3132, Adjusted R-squared:  0.3098
## F-statistic: 91.16 on 21 and 4198 DF, p-value: < 2.2e-16
```

```
library(gtsummary)
```

```
## #BlackLivesMatter
```

```
final_model1 <- model$finalModel
final_model2 <- model2$finalModel
final_model3 <- model3$finalModel

tbl1 <- tbl_regression(final_model1, exponentiate = FALSE) %>%
  modify_header(label = "***Model 1: avg_processing**")

tbl2 <- tbl_regression(final_model2, exponentiate = FALSE) %>%
  modify_header(label = "***Model 2: avg_ISS_processing**")

tbl3 <- tbl_regression(final_model3, exponentiate = FALSE) %>%
  modify_header(label = "***Model 3: avg_ABN_processing**")

tbl_merged <- tbl_merge(
  list(tbl1, tbl2, tbl3),
  tab_spanner = c("Average Processing", "Issued Processing", "Abandoned Processing")
```

```

)

# Convert tbl_merged to a gt table first
gt_table <- tbl_merged %>% as_gt()

# Save the gt table as HTML
gt::gtsave(gt_table, filename = "merged_table_lm.html")

predictions <- predict(model, newdata = test_data_lm)
predictions2 <- predict(model2, newdata = test_data_lm)
predictions3 <- predict(model3, newdata = test_data_lm)

library(Metrics)

##
## Attaching package: 'Metrics'

## The following objects are masked from 'package:caret':
##
##   precision, recall

rmse_avg_processing <- rmse(test_data_lm$avg_processing, predictions)
print(paste("RMSE for avg_processing:", rmse_avg_processing))

## [1] "RMSE for avg_processing: 303.372131914418"

rmse_avg_ISS_processing <- rmse(test_data_lm$avg_ISS_processing, predictions2)
print(paste("RMSE for avg_ISS_processing:", rmse_avg_ISS_processing))

## [1] "RMSE for avg_ISS_processing: 360.455134495717"

rmse_avg_ABN_processing <- rmse(test_data_lm$avg_ABN_processing, predictions3)
print(paste("RMSE for avg_ABN_processing:", rmse_avg_ABN_processing))

## [1] "RMSE for avg_ABN_processing: 326.32728593549"

```

Insights :

- What are the organizational and social factors associated with examiner attrition

```

train_control <- trainControl(method = "cv", number = 10, classProbs = TRUE, summaryFunction = twoClassSummary)

attrition_model <- train(separation_indicator ~ tenure_days + ISSUED_applications +
                        avg_ISS_processing + avg_ABN_processing + au_moves +
                        data = train_data,
                        method = "glm",
                        family = "binomial",
                        trControl = train_control,
                        preProcess = "scale",
                        metric = "ROC")

```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
attrition_model2 <- train(separation_indicator ~ tenure_days + ISSUED_applications
                          + avg_ISS_processing + avg_ABN_processing + au_moves +
                          data = train_data,
                          method = "glm",
                          family = "binomial",
                          trControl = train_control,
                          preProcess = "scale",
                          metric = "ROC")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
attrition_model3 <- train(separation_indicator ~ tenure_days + ISSUED_applications
  avg_ISS_processing + avg_ABN_processing + au_moves +
  data = train_data,
  method = "glm",
  family = "binomial",
  trControl = train_control,
  preProcess = "scale",
  metric = "ROC")
```

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

```
attrition_model4 <- train(separation_indicator ~ tenure_days + ISSUED_applications + abandoned_applicat
  avg_ISS_processing + avg_ABN_processing + au_moves + +
  gender + race + gender * tenure_days + race * tenure_days,
  data = train_data,
  method = "glm",
  family = "binomial",
  trControl = train_control,
  preProcess = "scale",
  metric = "ROC")
```

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =
10, : These variables have zero variances: raceother, tenure_days:raceother

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :  
## prediction from rank-deficient fit; attr(*, "non-estim") has doubtful cases
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :  
## prediction from rank-deficient fit; attr(*, "non-estim") has doubtful cases
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
final_model1 <- attrition_model$finalModel  
final_model2 <- attrition_model2$finalModel  
final_model3 <- attrition_model3$finalModel  
final_model4 <- attrition_model4$finalModel
```

```
library(gtsummary)
```

```
tbl1 <- tbl_regression(final_model1)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

tbl4 <- tbl_regression(final_model4)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]


```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
tbl_merged <- tbl_merge(
  list(tbl1, tbl2, tbl3, tbl4),
  tab_spanner = c("Model 1", "Model 2", "Model 3", "Model 4")
)

tbl_merged
```

```
## Table printed with 'knitr::kable()', not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include 'message = FALSE' in code chunk header.
```

Characteristic	log(OR)	95% CI	p-value	log(OR)	95% CI	p-value	log(OR)	95% CI	p-value	log(OR)	95% CI	p-value
tenure_days	-0.25	-0.43, -0.08	0.004	-0.21	-0.45, 0.03	0.084	-0.18	-0.41, 0.04	0.11	-0.27	-0.53, -0.01	0.040
ISSUED_applications	6.0	6.0, 7.8	<0.001	6.9	6.0, 7.8	<0.001	6.9	6.0, 7.8	<0.001	-0.80	-0.94, -0.67	<0.001
abandoned_applications	3.1	3.1, 4.2	<0.001	3.6	3.1, 4.2	<0.001	3.7	3.1, 4.2	<0.001	-0.66	-0.78, -0.55	<0.001
avg_processing	0.37	0.15, 0.60	0.001	0.37	0.15, 0.61	0.001	0.37	0.14, 0.60	0.002	1.7	1.5, 1.9	<0.001
avg_ISS_processing	0.03	0.03, 0.40	0.021	0.22	0.03, 0.40	0.020	0.23	0.04, 0.41	0.017	-0.23	-0.40, -0.05	0.012
avg_ABN_processing	0.58	0.58, -0.12	0.003	-0.35	-0.59, -0.13	0.003	-0.35	-0.58, -0.12	0.003	-1.4	-1.6, -1.2	<0.001
au_moves	0.01	-0.13, 0.14	>0.9	0.01	-0.13, 0.14	>0.9	0.01	-0.13, 0.14	>0.9	0.22	0.11, 0.33	<0.001
TC_1700_count	8.9	-10, -7.9	<0.001	-8.9	-10, -7.9	<0.001	-9.0	-10, -7.9	<0.001			
TC_1600_count	8.4	-9.4, -7.4	<0.001	-8.4	-9.4, -7.4	<0.001	-8.4	-9.4, -7.5	<0.001			
TC_2100_count	5.7	-6.4, -5.1	<0.001	-5.7	-6.4, -5.1	<0.001	-5.8	-6.5, -5.1	<0.001			

Characteristic	log(OR)	95% CI	p-value	log(OR)	95% CI	p-value	log(OR)	95% CI	p-value	log(OR)	95% CI	p-value
TC_2400_count	5.8	-6.4, -5.2	<0.001	-5.8	-6.4, -5.2	<0.001	-5.8	-6.4, -5.2	<0.001			
gendermale	0.03	-0.07, 0.13	0.5	0.15	-0.17, 0.46	0.3	0.04	-0.06, 0.14	0.5	0.29	-0.01, 0.58	0.054
genderunknown	0.04	-0.06, 0.14	0.4	-0.01	-0.31, 0.29	>0.9	0.04	-0.06, 0.14	0.5	-0.03	-0.32, 0.25	0.8
raceblack	0.00	-0.08, 0.09	>0.9	0.00	-0.08, 0.09	>0.9	0.14	-0.14, 0.45	0.3	0.12	-0.14, 0.40	0.4
raceHispanic	0.01	-0.07, 0.10	0.8	0.01	-0.08, 0.10	0.8	-0.08	-0.31, 0.17	0.5	-0.10	-0.33, 0.14	0.4
raceother	-0.01	-0.09, 0.08	0.8	-0.01	-0.09, 0.08	0.9	-7.9	-125, 30	>0.9	-8.1	-119, 28	>0.9
racewhite	-0.05	-0.14, 0.05	0.3	-0.05	-0.14, 0.05	0.4	0.11	-0.18, 0.38	0.5	-0.09	-0.36, 0.18	0.5
tenure_days:gendermale				-0.13	-0.46, 0.20	0.4				-0.28	-0.59, 0.02	0.069
tenure_days:genderunknown				0.06	-0.23, 0.34	0.7				0.01	-0.27, 0.28	>0.9
tenure_days:raceblack							-0.14	-0.44, 0.12	0.3	-0.16	-0.42, 0.09	0.2
tenure_days:raceHispanic							0.10	-0.14, 0.34	0.4	0.13	-0.10, 0.35	0.3
tenure_days:raceother							8.7	-17, 140	>0.9	8.9	-15, 134	>0.9
tenure_days:racewhite							-0.18	-0.49, 0.13	0.3	0.10	-0.20, 0.39	0.5

```
summary(attrition_model)
```

```
##
## Call:
## NULL
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    3.3055241  0.2391116  13.824 < 2e-16 ***
## tenure_days   -0.2536503  0.0882456  -2.874  0.00405 **
## ISSUED_applications  6.8766252  0.4476453  15.362 < 2e-16 ***
## abandoned_applications 3.6345260  0.2735496  13.287 < 2e-16 ***
## avg_processing    0.3715029  0.1166710   3.184  0.00145 **
## avg_ISS_processing  0.2182247  0.0946851   2.305  0.02118 *
## avg_ABN_processing -0.3465086  0.1168949  -2.964  0.00303 **
## au_moves         0.0076568  0.0679319   0.113  0.91026
## TC_1700_count   -8.9140335  0.5297911 -16.826 < 2e-16 ***
## TC_1600_count   -8.3965484  0.4904381 -17.121 < 2e-16 ***
## TC_2100_count   -5.7423927  0.3403175 -16.874 < 2e-16 ***
## TC_2400_count   -5.7866080  0.3086803 -18.746 < 2e-16 ***
## gendermale      0.0335379  0.0513045   0.654  0.51330
## genderunknown    0.0387226  0.0510863   0.758  0.44846
## raceblack       0.0006805  0.0437172   0.016  0.98758
```

```
## raceHispanic          0.0128345  0.0453060   0.283  0.77696
## raceother             -0.0078457  0.0385132  -0.204  0.83858
## racewhite             -0.0470811  0.0491524  -0.958  0.33813
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 5349.6  on 4220  degrees of freedom
## Residual deviance: 3314.8  on 4203  degrees of freedom
## AIC: 3350.8
##
## Number of Fisher Scoring iterations: 7
```

```
# Convert tbl_merged to a gt table first
gt_table <- tbl_merged %>% as_gt()

# Save the gt table as HTML
gt::gtsave(gt_table, filename = "merged_table_attrition_model.html")
```

```
predicted_probabilities1 <- predict(attrition_model, newdata = test_data, type = "prob")
predicted_probabilities2 <- predict(attrition_model2, newdata = test_data, type = "prob")
predicted_probabilities3 <- predict(attrition_model3, newdata = test_data, type = "prob")
predicted_probabilities4 <- predict(attrition_model4, newdata = test_data, type = "prob")
```

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following object is masked from 'package:Metrics':
##
##      auc
```

```
## The following objects are masked from 'package:stats':
##
##      cov, smooth, var
```

```
# Assuming the positive class is labeled as "1"
roc_curve1 <- roc(response = test_data$separation_indicator,
                  predictor = predicted_probabilities1[, "X1"])
```

```
## Setting levels: control = X0, case = X1
```

```
## Setting direction: controls < cases
```

```
auc_value1 <- auc(roc_curve1)
roc_curve2 <- roc(response = test_data$separation_indicator,
                  predictor = predicted_probabilities2[, "X1"])
```



```
## Setting levels: control = X0, case = X1
## Setting direction: controls < cases
```

```
auc_value2 <- auc(roc_curve2)
roc_curve3 <- roc(response = test_data$separation_indicator,
                  predictor = predicted_probabilities3[, "X1"])
```

```
## Setting levels: control = X0, case = X1
## Setting direction: controls < cases
```

```
auc_value3 <- auc(roc_curve3)
roc_curve4 <- roc(response = test_data$separation_indicator,
                  predictor = predicted_probabilities4[, "X1"])
```

```
## Setting levels: control = X0, case = X1
## Setting direction: controls < cases
```

```
auc_value4 <- auc(roc_curve4)

print(paste("AUC Model1:", auc_value1))
```

```
## [1] "AUC Model1: 0.878738184508336"
```

```
print(paste("AUC Model2:", auc_value2))
```

```
## [1] "AUC Model2: 0.878944318600854"
```

```
print(paste("AUC Model3:", auc_value3))
```

```
## [1] "AUC Model3: 0.876012633729492"
```

```
print(paste("AUC Model4:", auc_value4))
```

```
## [1] "AUC Model4: 0.815239722497704"
```