Andrew Singh and Mike Johnson
Machine Learning Project 2

Canterra Employee Attrition Findings and Recommendations

The Business Problem and Context

As external consultants we are supporting our client, Canterra, with data-driven solutions to improve related challenges with managing their current talent pool due to an attrition rate of 15% per year (affecting approximately 600 employees) leaving the company each year. For Canterra, one of their goals is to replace lost talent with new talent each year. Leadership believes their level of attrition is attributed to the following three organizational factors: 1) prior project timelines set by past employees are not maintained which damages Canterra's existing relationship with their consumers and partners, 2) adequately sized departments have to be maintained with new talent, and 3) incoming talent have to be trained to learn about their new roles and accompanying company policies.

Our aim in this report is to put forth a set of recommendations that shed light on the leadership's initial hypothesis of the relationship between higher job satisfaction and higher number of total working years reducing employee attrition, while at the same time supporting an in-depth analysis of the marketing department's ask to understand how demographic factors affect employee attrition.

Executive Summary and Recommendations

The analysis included in this report identified **business travel (frequent or rare)** and **marital status of single** significantly increase the probability of attrition. Other predictors that increase the probability of attrition include **number of companies worked** and **years at company**.

On the other hand, **job satisfaction (very high, high, or medium)** significantly reduces the probability of attrition. In addition, **total working years** reduced the probability of attrition.

Our recommendations based on these predictors are:

- **Provide additional support for employees who need to travel.** This includes improved travel allowances and investing in wellness programs to help alleviate the stress associated with travel. In addition, consider implementing technology solutions such as video conferencing to reduce the need for travel.

- **Develop targeted employee engagement strategies for single employees to help them feel more connected to the organization**. This could include social events, employee mentorship programs, and community-building initiatives.
- **Conduct regular employee satisfaction surveys to identify pain points and areas for improvement**. Invest in programs that improve job satisfaction such as employee recognition, career development opportunities, and initiatives to improve employee work-life balance.
- **Develop programs to incentivize long-term commitment to the organization.** This includes offering clear career progression or promotional paths, tailored professional development plans, and retention bonuses.

By implementing these strategies, Canterra can improve employee retention and build a more committed workforce to their organization.

Predictive Model

Our team chose to use a logistic predictive model to forecast employee attrition instead of classical predictive methods. Given that the management team is interested in a binary outcome (attrition vs. no attrition), a logistic predictive model is ideal for this analysis. Unlike numerical outcomes, our focus is on categorical choices, making logistic regression more suitable.

Using logistic prediction models helps mitigate the issue of multicollinearity, which refers to how other independent variables can affect or confound the outcomes we're interested in. This approach allows us to observe the true differences between the two outcomes, enabling us to fine-tune recommendations for improving recruitment strategies.

By clearly understanding these dynamics, we can provide precise and actionable insights to enhance talent retention.

Model Interpretation

The final model used had an **Area Under Curve (AUC)** value of 0.76 - suggesting that it is able to predict attrition with a good level of accuracy. Independent variables used in the final model include Age, Business Travel, Marital Status, Number of Companies Worked, Job Satisfaction, Total Working Years, Years at Company (squared).

The table below summarizes the average marginal effect to the probability of attrition. Statistically significant ($p < 0.05$) predictors have been color coded with **red** representing increased probability of attrition and **blue** representing decreasing probability of attrition.

Variable	Average Marginal Effect
Age	-0.003
Business Travel: Frequently	**+0.339**
Business Travel: Rarely	**+0.207**
Job Satisfaction: Very High	**-0.197**
Job Satisfaction: High	**-0.121**
Job Satisfaction: Medium	**-0.098**
Marital Status: Married	+0.060
Marital Status: Single	**+0.217**
Number of Companies Worked	**+0.031**
Total Working Years	**-0.022**
Years at Company (Squared)	**+0.007**

See Figure 1 in the Appendix for the ROC curve and other model metrics.

Key takeaways from the final model include:

- Business Travel (Frequently or Rarely), Marital Status (Single), number of companies worked, and years at the company increase the probability of attrition.
- Job Satisfaction (Very High, High, or Medium) and Total Working Years decrease the probability of attrition.

Distribution Estimates

For our analysis, the concept of a marginal distribution was leveraged to explain our findings, which is the unconditional probability in a cross-tab, where the probability of an event with no

strings attached or no conditions. In relation to understanding employee attrition, the marginal distribution allows the leadership team to comprehend how one category whether higher job satisfaction or higher number of total working years might reduce attrition conditioned by demographic variables such as gender, education, and age play a role as well.

The estimated marginal distribution of attrition without the influence of the logistic prediction model is 84% of employees were not affected by attrition, while 16% of employees were affected by attrition.

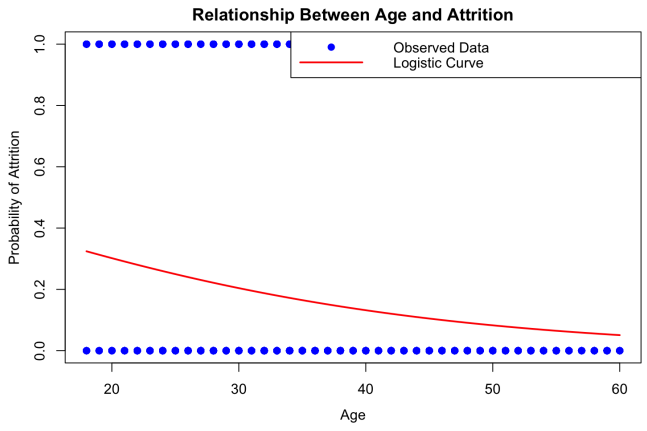
Findings from Exploratory Data Analysis

When looking at the various ages of those affected by attrition, those that are younger under the age of 25 and those that are older over the age of 55. The conditional distribution of attrition for these two age groups is 23% for younger employees, relative to older employees the attrition rate was approximately 14%.

	Age Statistics				
Attrition	**Mean**	**Median**	**Standard Deviation**	**Minimum**	**Maximum**
No	37.6	36	8.9	18	60
Yes	33.7	32	9.7	18	58

General Findings for the Marketing Leadership's Inquiry

In performing analysis to answer the marketing department's inquiry about the relationship between age and attrition, generally the finding is that as age increases, the probability of attrition decreases. The graphical logistic predictive model demonstrates the relationship between age and attrition.



General Findings from Demographic Variables and Attrition Models

The marketing management team expressed interest in understanding how several demographic variables affected attrition. From the variety of models performed, general findings are included below.

Gender Distribution and Attribution (See Appendix: Figure 2)

When it comes to attrition rates, males outpaced females in both categories when it comes to staying with the company or leaving the company. As a result, gender was concluded not to be a significant predictor of attrition.

Age Distribution and Attrition (See Appendix: Figure 3)

In terms of age, employees are more likely to become displaced or stay with the company by the average of 30 years old. After the age 30, the rate of attrition drops incrementally by each decade relative to the start of someone's career around the age of 20 years old.

Education Distribution and Attrition (See Appendix: Figure 4)

In terms of education distribution, employees are more likely to become affected by attrition when they have a bachelors or masters degree, relative to those who do not have a degree associated with higher education. Employees who hold a terminal degree such as a doctorate were the least likely to leave their company.

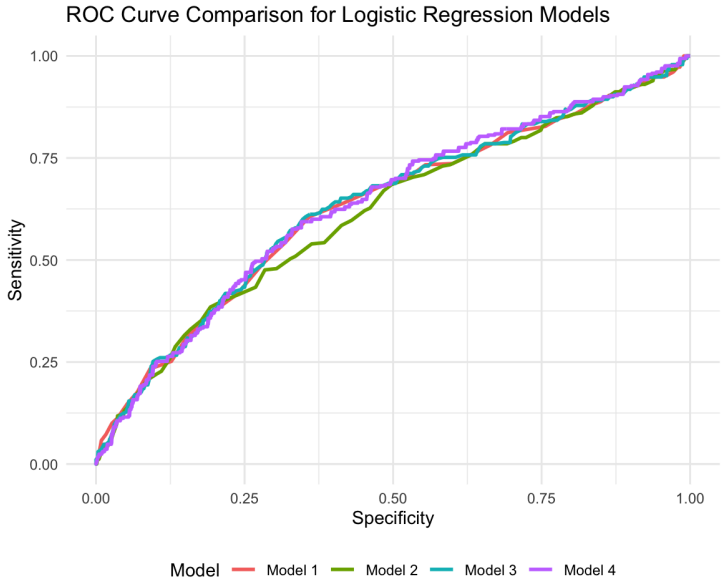
Models Supporting The Relationship between Demographics Factors and Attrition

The table below summarizes the performance of each predictive model, when comparing attrition to various factors in the dataset to see what the intended effects are, we find that model 3 performed the best with the lowest Akaike Information Criterion (AIC). All models performed similarly for Area Under Curve (AUC), Precision, and Recall.

	Variables	AIC	AUC	Precision	Recall
Model 1	Age	392.64	0.63	0.89	0.52
Model 2	Age, Gender	391.98	0.62	0.89	0.53

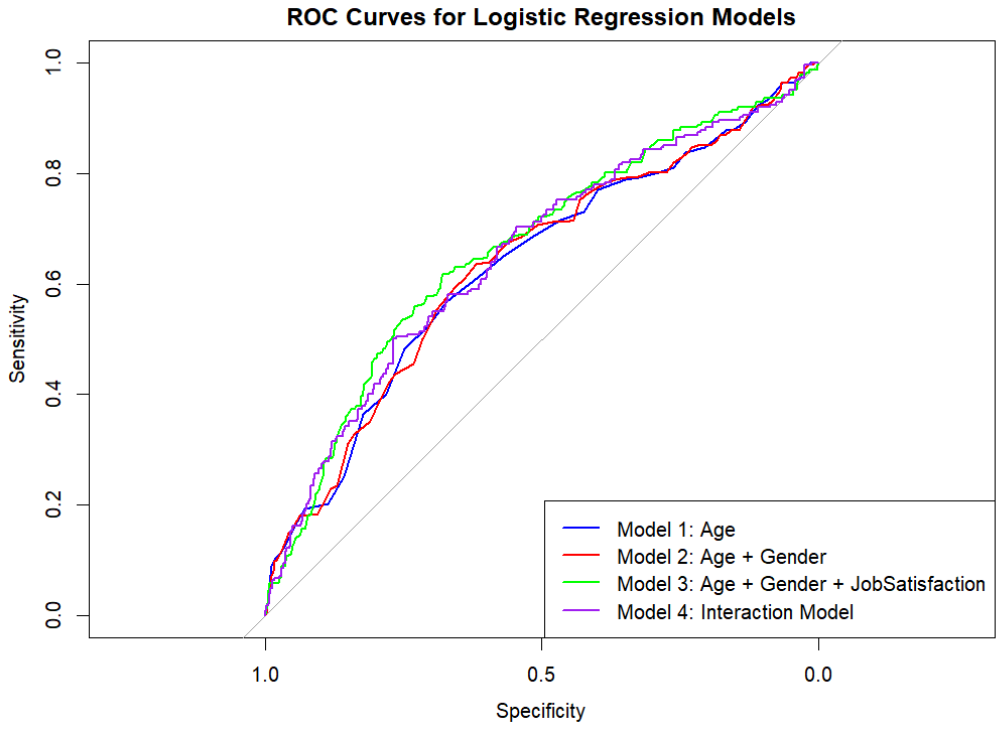
Model 3	Age, Gender, Job Satisfaction	387.91	0.63	0.90	0.59
Model 4	Age, Gender, Job Satisfaction, Income, Gender:Income	389.47	0.64	0.89	0.59

Similarly when plotting all the models graphically, each model is nearly identical to one another, with the exception of model 2 decreasing in its sensitivity.



Applying these models to a test set of data showed similar results. A summary of the test set is included below.

	Variables	**AIC**	**AUC**	**Precision**	**Recall**
Model 1	Age	393.99	0.63	0.89	0.57
Model 2	Age, Gender	395.89	0.64	0.89	0.60
Model 3	Age, Gender, Job Satisfaction	391.37	0.66	0.90	0.63
Model 4	Age, Gender, Job Satisfaction, Income, Gender:Income	393.35	0.65	0.88	0.61



Based on the model performance versus the test set, model 3 is the best performing model with the low AIC.

Appendix

Data Transformations

Recode Variables

Initially before our analysis, we performed pre-data processing to recode several variables to become categorical variables, as we wanted to represent underlying categories in detail. The re-coded variables are business travel, education, gender, job level, marital status, and satisfaction for both their job and environment. There were only two variables transformed into an integer datatype, which were the number of companies an employee worked at and their total number of working years.

Remove Unnecessary Variables

There were several variables that did not add any value to our analysis. As a result, employee ID and standard hours were removed. These columns were not found to be significant contributors to discovering a relationship with employee attrition or providing valuable insights or leadership recommendations.

Removing Null Values

Within the dataset, there were minimal blank cells with no information inputted for several variables such as number of companies worked, total working years, environment satisfaction, and job satisfaction. Due to the small number of missing values in the dataset, records with missing values were omitted.

Figure 1: Final Model Metrics & ROC Curve

AIC	AUC	Precision	Recall
355.09	0.74	0.92	0.65

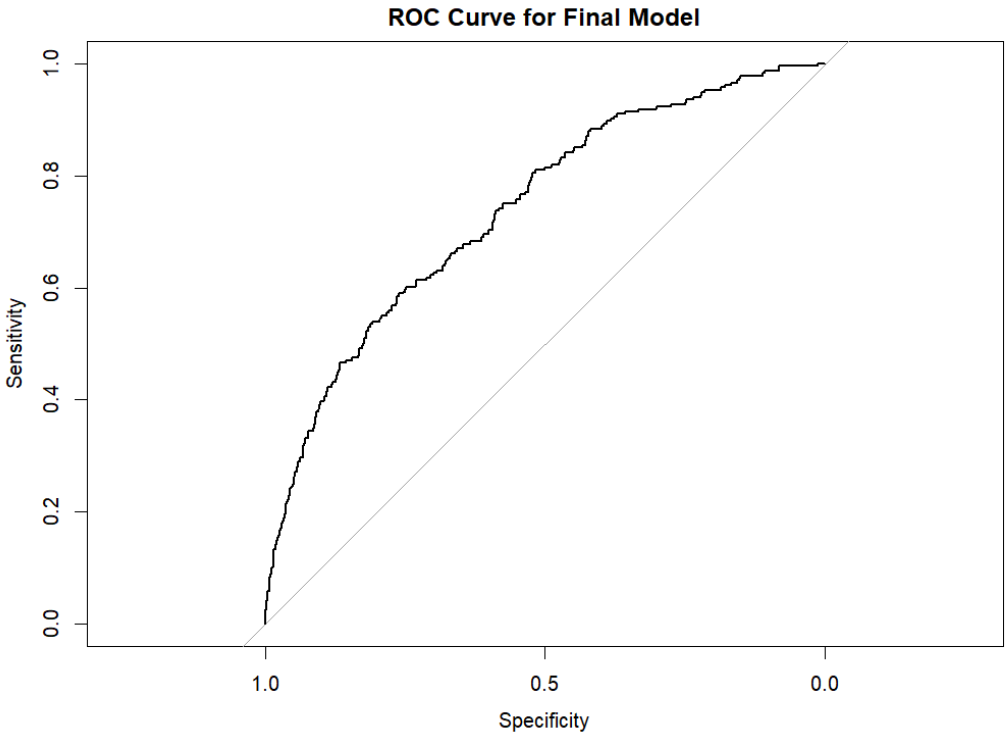


Figure 2: Gender Distribution

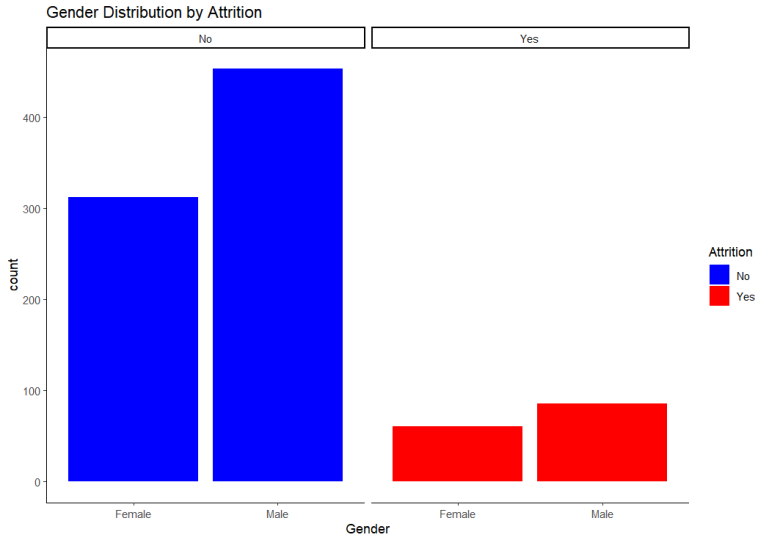


Figure 3: Age Distribution

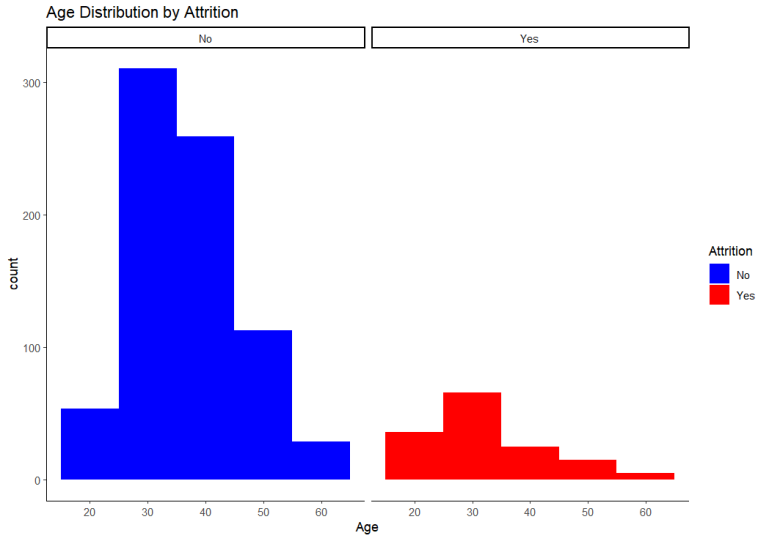
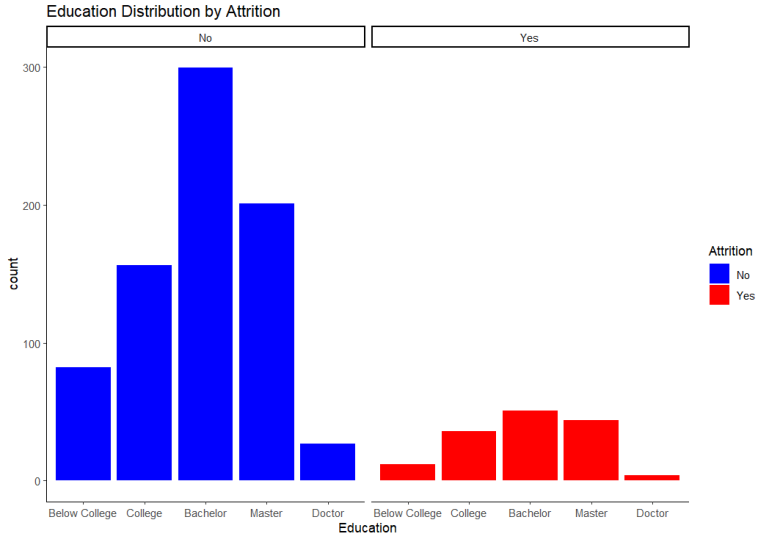
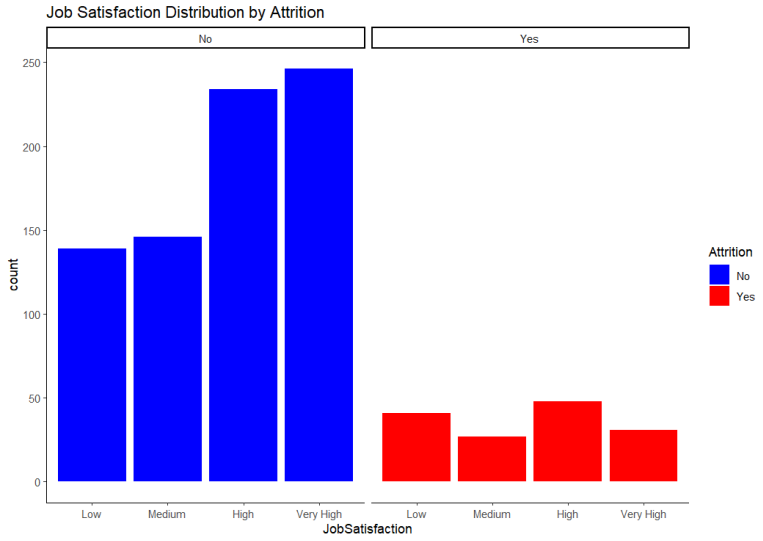
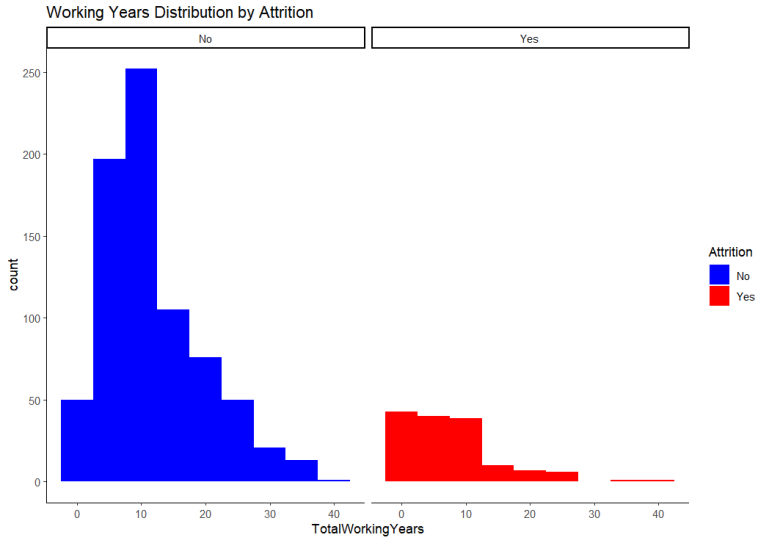


Figure 4: Education Distribution




Job Satisfaction Distribution by Attrition
No
Yes
250
200
150
100
50
0
count
Low
Medium
High
Very High
JobSatisfaction
Attrition
No
Yes


Working Years Distribution by Attrition
No
Yes
250
200
150
100
50
0
count
0
10
20
30
40
TotalWorkingYears
Attrition
No
Yes

Andrew Singh and Mike Johnson
Machine Learning Project 2

Technical Appendix

OPAN 6602 - Project 1

Mike Johnson / Andrew Singh - SAXA

Set up ----

Libraries

library(tidyverse)

library(caret)

library(GGally)

library(broom)

library(car) # Variance inflation factor

library(readxl) # read excel files

library(pROC) #Sampling-over and under, ROC and AUC curve

library(margins) # for marginal effects

Set random seed for reproducibility

set.seed(206)

Set viz theme

theme_set(theme_classic())

```
### Load Data ----

df = read_excel("Employee_Data_Project.xlsx")




table(df$Age)



ggplot(df, aes(Age)) +
  geom_histogram()
# Data structure
str(df)


# Update data types
df =
  df %>%
  mutate(
    # Dependent Variable
    Attrition = factor(Attrition),


    # Predictors
```

```
    BusinessTravel = factor(BusinessTravel),
    Education = factor(Education, levels = 1:5, labels = c("Below College", "College", "Bachelor", "Master", "Doctor")),
    Gender = factor(Gender),
    JobLevel = factor(JobLevel),
    MaritalStatus = factor(MaritalStatus),
    NumCompaniesWorked = as.numeric(NumCompaniesWorked),
    TotalWorkingYears = as.numeric(TotalWorkingYears),
    EnvironmentSatisfaction = factor(EnvironmentSatisfaction, levels = 1:4, labels = c("Low", "Medium", "High", "Very High")),
    JobSatisfaction = factor(JobSatisfaction, levels = 1:4, labels = c("Low", "Medium", "High", "Very High")))


# Remove Irrelevant Columns
df =
  df %>%
  select(
    -EmployeeID,
    -StandardHours)


# Check for NA's
na_summary = df %>%
  summarise_all(~ sum(is.na(.))) %>%
```

```
  pivot_longer(cols = everything(),
               names_to = "variable",
               values_to = "na_count") %>%
  filter(na_count > 0)


# How should we handle NAs?
na_summary


# Drop NA values
df = na.omit(df)


### Step 1: Create a train/test split ----


# Divide 30% of data to test set
test_indices = createDataPartition(1:nrow(df),
                                   times = 1,
                                   p = 0.3)


# Create training set
df_train = df[-test_indices[[1]], ]


# Create test set
```

```
df_test = df[test_indices[[1]], ]


# Create validation set

validation_indices = createDataPartition(1:nrow(df_train),

                                         times = 1,

                                         p = 0.3)


df_validation = df_train[-validation_indices[[1]],]


df_train = df_train[validation_indices[[1]],]


### Step 2: Data Exploration ----


# Summary of training set

summary(df_train)


#df_train %>%

# ggpairs(aes(color = Attrition, alpha = 0.4))


# Viz of attrition distribution

# Imbalanced classes. Will need to downsample.

df_train %>%
```

```
  ggplot(aes(x = Attrition)) +
  geom_bar(fill = "steelblue") +
  labs(title = "Attrition Distribution")

# Viz of relationship between Education and Attrition
df_train %>%
  ggplot(aes(x = Gender, fill = Attrition)) +
  geom_bar() +
  facet_grid(~Attrition) +
  labs(title = "Gender Distribution by Attrition")

# Viz of relationship between Age and Attrition
df_train %>%
  ggplot(aes(x = Age, fill = Attrition)) +
  geom_histogram(binwidth = 5, position = "dodge") +
  facet_grid(~Attrition) +
  labs(title = "Age Distribution by Attrition")

df_train %>%
  mutate(age_t = log(Age)) %>%
  ggplot(aes(x = age_t, fill = Attrition)) +
  geom_histogram() +
```

```
  facet_grid(~Attrition) +
  labs(title = "Age Transformed Distribution by Attrition")



# Viz of relationship between Education and Attrition
df_train %>%
  ggplot(aes(x = Education, fill = Attrition)) +
  geom_bar() +
  facet_grid(~Attrition) +
  labs(title = "Education Distribution by Attrition")

# Viz of relationship between Job Satisfaction and Attrition
df_train %>%
  ggplot(aes(x = JobSatisfaction, fill = Attrition)) +
  geom_bar() +
  facet_grid(~Attrition) +
  labs(title = "Job Satisfaction Distribution by Attrition")

# Viz of relationship between Working Years and Attrition
df_train %>%
  ggplot(aes(x = TotalWorkingYears, fill = Attrition)) +
  geom_histogram(binwidth = 5, position = "dodge") +
```

```
  facet_grid(~Attrition) +
  labs(title = "Working Years Distribution by Attrition")


df_train %>%
  mutate(workingyears_t = log(TotalWorkingYears)) %>%
  ggplot(aes(x = workingyears_t, fill = Attrition)) +
  geom_histogram() +
  facet_grid(~Attrition) +
  labs(title = "Working Years Transformed Distribution by Attrition")



### Step 3: Data pre-processing ----


# Downsampling
downsample_df = downSample(x = df_train[ , colnames(df_train) != "Attrition"],
                           y = df_train$Attrition)


colnames(downsample_df)[ncol(downsample_df)] = "Attrition"


downsample_df %>%
  ggplot(aes(x = Attrition)) +
  geom_bar(fill = "steelblue") +
```

```
  labs(title = "Attrition Distribution")


### Step 4: Feature Engineering ----



### Step 5: Feature & Model Selection ----


# Initial Model
f1 = glm(
  Attrition ~ . +
    I(Age ^ 2) +
    I(DistanceFromHome ^ 2) +
    I(Income ^ 2) +
    I(NumCompaniesWorked ^ 2) +
    I(TotalWorkingYears ^ 2) +
    I(TrainingTimesLastYear ^ 2) +
    I(YearsAtCompany ^ 2) +
    I(YearsWithCurrManager ^ 2),
  data = downsample_df,
  family = binomial("logit"))


summary(f1)
```

```
vif(f1)


roc1 = roc(
  data =
    tibble(
      actual =
        df_train %>%  # not using balanced data for evaluation
        select(Attrition) %>%
        unlist(),
        predicted = predict(f1, df_train)),
  "actual",
  "predicted"
)


plot(roc1)


roc1$auc


# Stepwise Regression
f_step = step(object = f1,
              direction = "both")
```

```
summary(f_step)


vif(f_step)


roc_step = roc(
  data =
    tibble(
      actual =
        df_train %>%  # not using balanced data for evaluation
        select(Attrition) %>%
        unlist(),
      predicted = predict(f_step, df_train)),
  "actual",
  "predicted"
)


plot(roc_step)


roc_step$auc

roc1$auc
```

```
# Final Model

f_final = glm(

  Attrition ~

    Age +

    BusinessTravel +

    MaritalStatus +

    NumCompaniesWorked +

    JobSatisfaction +

    TotalWorkingYears +

    I(YearsAtCompany^2),

  data = downsample_df,

  family = binomial("logit"))


summary(f_final)


vif(f_final)


roc_final = roc(

  data =

    tibble(

      actual =

        df_train %>%  # not using balanced data for evaluation
```

```
        select(Attrition) %>%
        unlist(),
      predicted = predict(f_final, df_train)),
  "actual",
  "predicted"
)


plot(roc_final)


roc_final$auc
roc_step$auc
roc1$auc


### Step 6: Model Validation ----


preds_validation = predict(f_final, df_validation)


roc_validation = roc(
  data =
    tibble(
      actual =
        df_validation %>%  # not using balanced data for evaluation
```

```
        select(Attrition) %>%
        unlist(),
      predicted = preds_validation),
  "actual",
  "predicted"
)

plot(roc_validation)

roc_final$auc
roc_validation$auc

### Step 7: Predictions and Conclusions ----

preds_test = predict(f_final, df_test)

roc_test = roc(
  data =
    tibble(
      actual =
        df_test %>%  # not using balanced data for evaluation
        select(Attrition) %>%
```

```
        unlist(),
      predicted = preds_test),
  "actual",
  "predicted"
)


plot(roc_final)


roc_final$auc
roc_test$auc


# Re-train the model on the whole data set for marginal effects/production


# Downsampling
downsample_prod = downSample(x = df[ , colnames(df) != "Attrition"],
                             y = df$Attrition)


colnames(downsample_prod)[ncol(downsample_prod)] = "Attrition"


# Production Model
f_prod = glm(
  Attrition ~
```

```
    Age +
    BusinessTravel +
    MaritalStatus +
    NumCompaniesWorked +
    JobSatisfaction +
    TotalWorkingYears +
    I(YearsAtCompany^2),
  data = downsample_prod,
  family = binomial("logit"))

summary(f_prod)

roc_prod = roc(
  data =
    tibble(
      actual =
        df %>%  # not using balanced data for evaluation
        select(Attrition) %>%
        unlist(),
      predicted = predict(f_prod, df)),
  "actual",
  "predicted"
```

```
)


plot(roc_prod)


roc_prod$auc


# Marginal Effects
coefs =
  tidy(f_prod) %>%
  mutate(odds = exp(estimate),
         odds_mfx = odds - 1)


coefs


mfx = margins(f_prod)


summary(mfx)


summary(f_prod)



### Model Comparison ----
```

```
# Create Models

model_1 = glm(Attrition ~ Age,
              data = downsample_df,
              family = binomial("logit"))


model_2 = glm(Attrition ~ Age + Gender,
              data = downsample_df,
              family = binomial("logit"))


model_3 = glm(Attrition ~ Age + Gender + JobSatisfaction,
              data = downsample_df,
              family = binomial("logit"))


model_4 = glm(Attrition ~ Age + Gender + JobSatisfaction + Income + Gender:Income,
              data = downsample_df,
              family = binomial("logit"))


##  Validate models


# Function to calculate model metrics
```

```
calc_metrics = function(model, data) {
  predicted_prob = predict(model, data, type = "response")
  predicted_class = ifelse(predicted_prob > 0.5, "Yes", "No")
  actual_values = data$Attrition

  auc = auc(roc(actual_values, predicted_prob))
  conf_matrix = confusionMatrix(factor(predicted_class,
                                       levels = c("No", "Yes")),
                                factor(actual_values, levels = c("No", "Yes")))
  precision = conf_matrix$byClass['Precision']
  recall = conf_matrix$byClass['Recall']

  return(list(AIC = AIC(model), AUC = auc, Precision = precision, Recall = recall))
}



# Create a table with all values
calc_metrics(model_1, df_validation)
calc_metrics(model_2, df_validation)
calc_metrics(model_3, df_validation)
calc_metrics(model_4, df_validation)
```

```
# Create a graph with all values
roc_data <- bind_rows(
  mutate(roc_1_validation, Model = "Model 1"),
  mutate(roc_2_validation, Model = "Model 2"),
  mutate(roc_3_validation, Model = "Model 3"),
  mutate(roc_4_validation, Model = "Model 4")
)

# Plot the ROC curves
ggplot(roc_data, aes(x = 1 - Specificity, y = Sensitivity, color = Model)) +
  geom_line(linewidth = 1) +
  labs(
    title = "ROC Curve Comparison for Logistic Regression Models",
    x = "Specificity",
    y = "Sensitivity",
    color = "Model"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")


# Actual Values
```

```
actual_values =
  df_validation %>%
  select(Attrition) %>%
  unlist()

# Model 1
model_1_validation = predict(model_1, df_validation)

roc_1_validation = roc(
  data =
    tibble(
      actual =
        df_validation %>%  # not using balanced data for evaluation
        select(Attrition) %>%
        unlist(),
      predicted = model_1_validation),
  "actual",
  "predicted"
)

plot(roc_1_validation)
```

```
roc_1_validation$auc


# Model 2

model_2_validation = predict(model_2, df_validation)


roc_2_validation = roc(
  data =
    tibble(
      actual =
        df_validation %>%  # not using balanced data for evaluation
        select(Attrition) %>%
        unlist(),
      predicted = model_2_validation),
  "actual",
  "predicted"
)


plot(roc_2_validation)


roc_2_validation$auc


# Model 3
```

```
model_3_validation = predict(model_3, df_validation)


roc_3_validation = roc(
  data =
    tibble(
      actual =
        df_validation %>%  # not using balanced data for evaluation
        select(Attrition) %>%
        unlist(),
      predicted = model_3_validation),
  "actual",
  "predicted"
)


plot(roc_3_validation)


roc_3_validation$auc


# Model 4
model_4_validation = predict(model_4, df_validation)


roc_4_validation = roc(
```

```
  data =
    tibble(
      actual =
        df_validation %>%  # not using balanced data for evaluation
        select(Attrition) %>%
        unlist(),
      predicted = model_4_validation),
  "actual",
  "predicted"
)

plot(roc_4_validation)

roc_4_validation$auc
```

Part 3.1

```
# Calculate the marginal distribution of Attrition
```

```
marginal_distribution <- prop.table(table(df$Attrition))


# Print the results

cat("Estimated Marginal Distribution of Attrition:\n")

print(marginal_distribution)


## Part 3.2



# Define the younger and older age groups

younger_age <- 25

older_age <- 55


# Filter the data for the specified age groups

younger_group <- subset(df, Age == younger_age)

older_group <- subset(df, Age == older_age)


# Calculate the attrition rate for each group

younger_attrition_rate <- sum(younger_group$Attrition == "Yes") / nrow(younger_group)

older_attrition_rate <- sum(older_group$Attrition == "Yes") / nrow(older_group)


# Replace NaN with 0 if there are no records for a group
```

```
younger_attrition_rate <- ifelse(is.nan(younger_attrition_rate), 0, younger_attrition_rate)

older_attrition_rate <- ifelse(is.nan(older_attrition_rate), 0, older_attrition_rate)


# Print the results

cat("Attrition rate for younger employees (age 25):", younger_attrition_rate, "\n")

cat("Attrition rate for older employees (age 55):", older_attrition_rate, "\n")



## 4.1


# Convert Attrition to a binary variable for logistic regression

df$AttritionBinary <- ifelse(df$Attrition == "Yes", 1, 0)


# Fit a logistic regression model

logit_model <- glm(AttritionBinary ~ Age, data = df, family = binomial)


# Create a sequence of ages for prediction

age_range <- seq(min(df$Age), max(df$Age), length.out = 100)


# Predict probabilities of attrition using the logistic model

predicted_probs <- predict(logit_model, newdata = data.frame(Age = age_range), type =
"response")
```

```
summary(predicted_probs)


# Plot the relationship
plot(df$Age, df$AttritionBinary,
     xlab = "Age",
     ylab = "Probability of Attrition",
     main = "Relationship Between Age and Attrition",
     pch = 19, col = "blue",
     xlim = c(min(age_range), max(age_range)),
     ylim = c(0, 1))


# Add the logistic regression curve
lines(age_range, predicted_probs, col = "red", lwd = 2)


# Add legend
legend("topright", legend = c("Observed Data", "Logistic Curve"),
       col = c("blue", "red"), pch = c(19, NA), lty = c(NA, 1), lwd = c(NA, 2))
```