Project - Phase 2 Deliverables

2024-10-12

# Load necessary libraries  
library(tidyverse)

## ── 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 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── 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(tidymodels)

## ── Attaching packages ────────────────────────────────────── tidymodels 1.2.0 ──  
## ✔ broom 1.0.6 ✔ rsample 1.2.1  
## ✔ dials 1.3.0 ✔ tune 1.2.1  
## ✔ infer 1.0.7 ✔ workflows 1.1.4  
## ✔ modeldata 1.4.0 ✔ workflowsets 1.1.0  
## ✔ parsnip 1.2.1 ✔ yardstick 1.3.1  
## ✔ recipes 1.1.0   
## ── Conflicts ───────────────────────────────────────── tidymodels\_conflicts() ──  
## ✖ scales::discard() masks purrr::discard()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ recipes::fixed() masks stringr::fixed()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ yardstick::spec() masks readr::spec()  
## ✖ recipes::step() masks stats::step()  
## • Dig deeper into tidy modeling with R at https://www.tmwr.org

library(VIM)

## Loading required package: colorspace  
## Loading required package: grid  
## VIM is ready to use.  
##   
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues  
##   
## Attaching package: 'VIM'  
##   
## The following object is masked from 'package:recipes':  
##   
## prepare  
##   
## The following object is masked from 'package:datasets':  
##   
## sleep

library(caret)

## Loading required package: lattice  
##   
## Attaching package: 'caret'  
##   
## The following objects are masked from 'package:yardstick':  
##   
## precision, recall, sensitivity, specificity  
##   
## The following object is masked from 'package:purrr':  
##   
## lift

library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(mice)

##   
## Attaching package: 'mice'  
##   
## The following object is masked from 'package:stats':  
##   
## filter  
##   
## The following objects are masked from 'package:base':  
##   
## cbind, rbind

library(yardstick)

train\_data = read\_csv("/Users/ellismacbookair/Library/Mobile Documents/com~apple~CloudDocs/Documents/UNCW MBA/BAN - 502 Pred Analytics/product-failure-kaggle-competition-fall-24/train.csv")

## Rows: 26570 Columns: 26  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (4): product\_code, attribute\_0, attribute\_1, failure  
## dbl (22): id, loading, attribute\_2, attribute\_3, measurement\_0, measurement\_...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

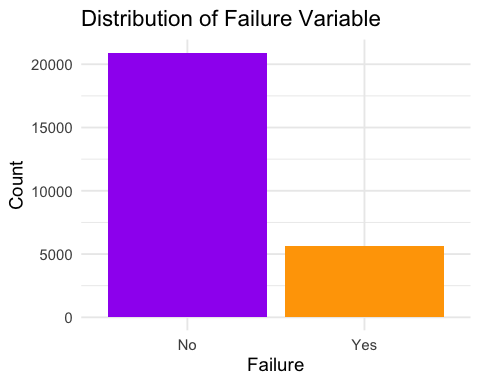
# Checking structure of the dataset  
str(train\_data)

## spc\_tbl\_ [26,570 × 26] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ id : num [1:26570] 0 1 2 3 4 5 6 7 8 9 ...  
## $ product\_code : chr [1:26570] "A" "A" "A" "A" ...  
## $ loading : num [1:26570] 80.1 84.9 82.4 101.1 188.1 ...  
## $ attribute\_0 : chr [1:26570] "material\_7" "material\_7" "material\_7" "material\_7" ...  
## $ attribute\_1 : chr [1:26570] "material\_8" "material\_8" "material\_8" "material\_8" ...  
## $ attribute\_2 : num [1:26570] 9 9 9 9 9 9 9 9 9 9 ...  
## $ attribute\_3 : num [1:26570] 5 5 5 5 5 5 5 5 5 5 ...  
## $ measurement\_0 : num [1:26570] 7 14 12 13 9 11 12 4 9 10 ...  
## $ measurement\_1 : num [1:26570] 8 3 1 2 2 4 2 8 6 4 ...  
## $ measurement\_2 : num [1:26570] 4 3 5 6 8 0 4 8 5 7 ...  
## $ measurement\_3 : num [1:26570] 18 18.2 18.1 17.3 19.3 ...  
## $ measurement\_4 : num [1:26570] 12.5 11.5 11.7 11.2 12.9 ...  
## $ measurement\_5 : num [1:26570] 15.7 17.7 16.7 18.6 17 ...  
## $ measurement\_6 : num [1:26570] 19.3 17.9 18.2 18.3 15.7 ...  
## $ measurement\_7 : num [1:26570] 11.7 12.7 12.7 12.6 11.3 ...  
## $ measurement\_8 : num [1:26570] 20.2 17.9 18.3 19.1 18.1 ...  
## $ measurement\_9 : num [1:26570] 10.7 12.4 12.7 12.5 10.3 ...  
## $ measurement\_10: num [1:26570] 15.9 17.9 15.6 16.3 17.1 ...  
## $ measurement\_11: num [1:26570] 17.6 17.9 NA 18.4 19.9 ...  
## $ measurement\_12: num [1:26570] 15.2 11.8 13.8 10 12.4 ...  
## $ measurement\_13: num [1:26570] 15 14.7 16.7 15.2 16.2 ...  
## $ measurement\_14: num [1:26570] NA 15.4 18.6 15.6 12.8 ...  
## $ measurement\_15: num [1:26570] 13 14.4 14.1 16.2 13.2 ...  
## $ measurement\_16: num [1:26570] 14.7 15.6 17.9 17.2 16.4 ...  
## $ measurement\_17: num [1:26570] 764 682 663 826 580 ...  
## $ failure : chr [1:26570] "No" "No" "No" "No" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. id = col\_double(),  
## .. product\_code = col\_character(),  
## .. loading = col\_double(),  
## .. attribute\_0 = col\_character(),  
## .. attribute\_1 = col\_character(),  
## .. attribute\_2 = col\_double(),  
## .. attribute\_3 = col\_double(),  
## .. measurement\_0 = col\_double(),  
## .. measurement\_1 = col\_double(),  
## .. measurement\_2 = col\_double(),  
## .. measurement\_3 = col\_double(),  
## .. measurement\_4 = col\_double(),  
## .. measurement\_5 = col\_double(),  
## .. measurement\_6 = col\_double(),  
## .. measurement\_7 = col\_double(),  
## .. measurement\_8 = col\_double(),  
## .. measurement\_9 = col\_double(),  
## .. measurement\_10 = col\_double(),  
## .. measurement\_11 = col\_double(),  
## .. measurement\_12 = col\_double(),  
## .. measurement\_13 = col\_double(),  
## .. measurement\_14 = col\_double(),  
## .. measurement\_15 = col\_double(),  
## .. measurement\_16 = col\_double(),  
## .. measurement\_17 = col\_double(),  
## .. failure = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

head(train\_data)

## # A tibble: 6 × 26  
## id product\_code loading attribute\_0 attribute\_1 attribute\_2 attribute\_3  
## <dbl> <chr> <dbl> <chr> <chr> <dbl> <dbl>  
## 1 0 A 80.1 material\_7 material\_8 9 5  
## 2 1 A 84.9 material\_7 material\_8 9 5  
## 3 2 A 82.4 material\_7 material\_8 9 5  
## 4 3 A 101. material\_7 material\_8 9 5  
## 5 4 A 188. material\_7 material\_8 9 5  
## 6 5 A 75.4 material\_7 material\_8 9 5  
## # ℹ 19 more variables: measurement\_0 <dbl>, measurement\_1 <dbl>,  
## # measurement\_2 <dbl>, measurement\_3 <dbl>, measurement\_4 <dbl>,  
## # measurement\_5 <dbl>, measurement\_6 <dbl>, measurement\_7 <dbl>,  
## # measurement\_8 <dbl>, measurement\_9 <dbl>, measurement\_10 <dbl>,  
## # measurement\_11 <dbl>, measurement\_12 <dbl>, measurement\_13 <dbl>,  
## # measurement\_14 <dbl>, measurement\_15 <dbl>, measurement\_16 <dbl>,  
## # measurement\_17 <dbl>, failure <chr>

# Create a bar plot for the failure variable  
ggplot(train\_data, aes(x = failure)) +  
 geom\_bar(fill = c("purple", "orange")) + # Colors for 'No' and 'Yes'  
 labs(title = "Distribution of Failure Variable",   
 x = "Failure",   
 y = "Count") + # Add labels for axes  
 theme\_minimal(base\_size = 14) # Use minimal theme for clean layout

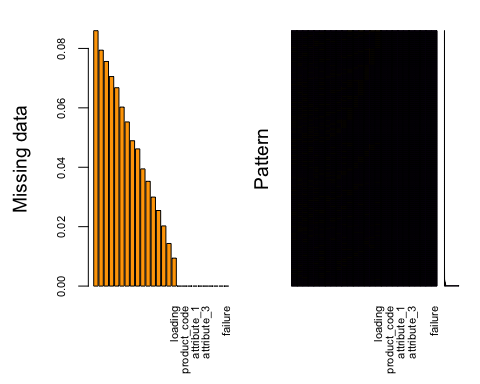


# Calculate the number of failures and non-failures  
failure\_counts = train\_data %>% count(failure)  
  
# Calculate percentages  
failure\_counts = failure\_counts %>%  
 mutate(percent = n / sum(n) \* 100)  
  
# Show the breakdown in console  
failure\_counts

## # A tibble: 2 × 3  
## failure n percent  
## <chr> <int> <dbl>  
## 1 No 20921 78.7  
## 2 Yes 5649 21.3

# Plot missing data pattern with VIM package  
aggr(train\_data, col=c('purple','orange'), numbers=TRUE, sortVars=TRUE, labels=names(train\_data), cex.axis=.7, gap=3, ylab=c("Missing data","Pattern"))

## Warning in plot.aggr(res, ...): not enough vertical space to display  
## frequencies (too many combinations)



##   
## Variables sorted by number of missings:   
## Variable Count  
## measurement\_17 0.085961611  
## measurement\_16 0.079412872  
## measurement\_15 0.075611592  
## measurement\_14 0.070530674  
## measurement\_13 0.066767030  
## measurement\_12 0.060255928  
## measurement\_11 0.055250282  
## measurement\_10 0.048927362  
## measurement\_9 0.046179902  
## measurement\_8 0.039442981  
## measurement\_7 0.035265337  
## measurement\_6 0.029958600  
## measurement\_5 0.025442228  
## measurement\_4 0.020248400  
## measurement\_3 0.014339481  
## loading 0.009409108  
## id 0.000000000  
## product\_code 0.000000000  
## attribute\_0 0.000000000  
## attribute\_1 0.000000000  
## attribute\_2 0.000000000  
## attribute\_3 0.000000000  
## measurement\_0 0.000000000  
## measurement\_1 0.000000000  
## measurement\_2 0.000000000  
## failure 0.000000000

# Perform missing data imputation using mice  
imputed\_data = mice(train\_data, m = 5, method = 'pmm', seed = 123)

##   
## iter imp variable  
## 1 1 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 1 2 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 1 3 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 1 4 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 1 5 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 2 1 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 2 2 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 2 3 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 2 4 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 2 5 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 3 1 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 3 2 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 3 3 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 3 4 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 3 5 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 4 1 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 4 2 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 4 3 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 4 4 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 4 5 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 5 1 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 5 2 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 5 3 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 5 4 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17  
## 5 5 loading measurement\_3 measurement\_4 measurement\_5 measurement\_6 measurement\_7 measurement\_8 measurement\_9 measurement\_10 measurement\_11 measurement\_12 measurement\_13 measurement\_14 measurement\_15 measurement\_16 measurement\_17

## Warning: Number of logged events: 4

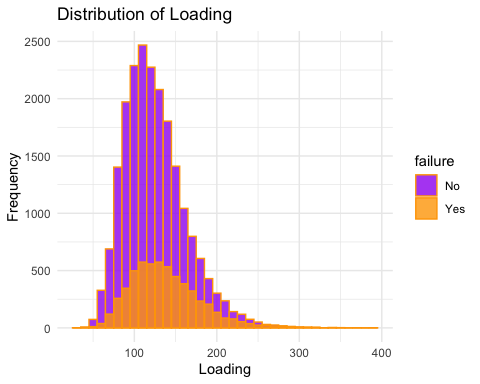
# Plot 2: Distribution of Loading by Failure  
ggplot(train\_data, aes(x = failure, y = loading, fill = failure)) +  
 geom\_boxplot() +  
 scale\_fill\_manual(values = c("No" = "purple", "Yes" = "orange")) +  
 labs(title = "Distribution of Loading by Failure", x = "Failure", y = "Loading") +  
 theme\_minimal()

## Warning: Removed 250 rows containing non-finite outside the scale range  
## (`stat\_boxplot()`).



# Plot the distribution of 'loading' with color fill based on 'failure' status  
ggplot(train\_data, aes(x = loading, fill = failure)) +  
 geom\_histogram(position = "identity", binwidth = 10, color = "orange", alpha = 0.8) +   
 scale\_fill\_manual(values = c("No" = "purple", "Yes" = "orange")) +  
 labs(title = "Distribution of Loading", x = "Loading", y = "Frequency") +  
 theme\_minimal()

## Warning: Removed 250 rows containing non-finite outside the scale range  
## (`stat\_bin()`).



# Convert all character columns to factors  
train\_data = train\_data %>%  
 mutate\_if(is.character, as.factor)

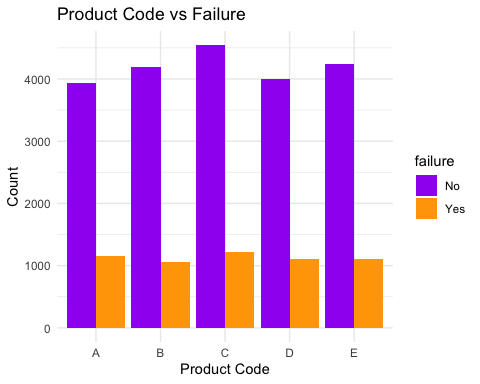
# Set up the logistic regression model  
log\_reg\_model = logistic\_reg() %>%   
 set\_engine("glm") %>%   
 set\_mode("classification")  
  
# Define the recipe (without pre-prepping)  
recipe = recipe(failure ~ loading + product\_code + attribute\_0 + attribute\_1, data = train\_data)  
  
# Create a workflow and add the model and recipe  
log\_reg\_workflow = workflow() %>%  
 add\_model(log\_reg\_model) %>%  
 add\_recipe(recipe)  
  
# Fit the model using the updated data  
log\_reg\_fit = fit(log\_reg\_workflow, data = train\_data)  
  
# Check the fit summary  
log\_reg\_fit

## ══ Workflow [trained] ══════════════════════════════════════════════════════════  
## Preprocessor: Recipe  
## Model: logistic\_reg()  
##   
## ── Preprocessor ────────────────────────────────────────────────────────────────  
## 0 Recipe Steps  
##   
## ── Model ───────────────────────────────────────────────────────────────────────  
##   
## Call: stats::glm(formula = ..y ~ ., family = stats::binomial, data = data)  
##   
## Coefficients:  
## (Intercept) loading product\_codeB   
## -2.227814 0.007631 -0.148892   
## product\_codeC product\_codeD product\_codeE   
## -0.086939 -0.053862 -0.107292   
## attribute\_0material\_7 attribute\_1material\_6 attribute\_1material\_8   
## NA NA NA   
##   
## Degrees of Freedom: 26319 Total (i.e. Null); 26314 Residual  
## (250 observations deleted due to missingness)  
## Null Deviance: 27260   
## Residual Deviance: 26830 AIC: 26840

# Define cross-validation folds  
cv\_folds = vfold\_cv(train\_data, v = 5)  
  
# Perform cross-validation  
log\_reg\_results = fit\_resamples(  
 log\_reg\_workflow,  
 resamples = cv\_folds,  
 metrics = metric\_set(accuracy, roc\_auc),  
 control = control\_resamples(save\_pred = TRUE)  
)  
  
# Collect metrics from cross-validation  
collect\_metrics(log\_reg\_results)

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.787 5 0.00168 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.588 5 0.00211 Preprocessor1\_Model1

# Bar chart for product code vs failure  
ggplot(train\_data, aes(x = product\_code, fill = failure)) +   
 geom\_bar(position = "dodge") +   
 scale\_fill\_manual(values = c("No" = "purple", "Yes" = "orange")) +   
 labs(title = "Product Code vs Failure", x = "Product Code", y = "Count") +  
 theme\_minimal()



# Bar chart for model performance comparison (optional)  
model\_performance = data.frame(  
 Model = rep(c("Logistic Regression", "Classification Tree", "Random Forest"), 2),  
 Metric = rep(c("Accuracy", "AUC"), each = 3),  
 Score = c(0.787, 0.687, 0.800, 0.588, 0.680, 0.850)  
)  
  
ggplot(model\_performance, aes(x = Model, y = Score, fill = Metric)) +   
 geom\_bar(stat = "identity", position = "dodge") +   
 scale\_fill\_manual(values = c("Accuracy" = "purple", "AUC" = "orange")) +  
 labs(title = "Model Performance Comparison", x = "Model", y = "Score") +  
 theme\_minimal()

