203 Final Project

2024-11-20

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(pROC)
```

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

##
## Attaching package: 'pROC'

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

```r
diabetes_data <- read.csv("diabetes_binary_5050split_health_indicators_BRFSS2015.csv", stringsAsFactors
```

1. Data Preparation

```r
str(diabetes_data)
```

```
## 'data.frame':    70692 obs. of  22 variables:
##  $ Diabetes_binary     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ HighBP              : num  1 1 0 1 0 0 0 0 0 0 ...
##  $ HighChol            : num  0 1 0 1 0 0 1 0 0 0 ...
##  $ CholCheck           : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ BMI                 : num  26 26 26 28 29 18 26 31 32 27 ...
##  $ Smoker              : num  0 1 0 1 1 0 1 1 0 1 ...
##  $ Stroke              : num  0 1 0 0 0 0 0 0 0 0 ...
##  $ HeartDiseaseorAttack: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ PhysActivity        : num  1 0 1 1 1 1 1 0 1 0 ...
##  $ Fruits              : num  0 1 1 1 1 1 1 1 1 1 ...
##  $ Veggies             : num  1 0 1 1 1 1 1 1 1 1 ...
##  $ HvyAlcoholConsump   : num  0 0 0 0 0 0 1 0 0 0 ...
##  $ AnyHealthcare       : num  1 1 1 1 1 0 1 1 1 1 ...
##  $ NoDocbcCost         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ GenHlth             : num  3 3 1 3 2 2 1 4 3 3 ...
##  $ MentHlth            : num  5 0 0 0 0 7 0 0 0 0 ...
##  $ PhysHlth            : num  30 0 10 3 0 0 0 0 0 6 ...
##  $ DiffWalk            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Sex                 : num  1 1 1 1 0 0 1 1 0 1 ...
##  $ Age                 : num  4 12 13 11 8 1 13 6 3 6 ...
##  $ Education           : num  6 6 6 6 5 4 5 4 6 4 ...
##  $ Income              : num  8 8 8 8 8 7 6 3 8 4 ...
```

```r
dim(diabetes_data)
```

```
## [1] 70692    22
```

2. Descriptive Statistics

```r
# Select only the key variables to summarize
key_variables <- diabetes_data %>% select(MentHlth, PhysHlth, Smoker, PhysActivity, Diabetes_binary)

# Function to check for outliers using IQR and count missing values
summary_stats <- apply(key_variables, 2, function(x) {
  q1 <- quantile(x, 0.25, na.rm = TRUE)  # First quartile
  q3 <- quantile(x, 0.75, na.rm = TRUE)  # Third quartile
  iqr <- q3 - q1  # Interquartile range
  lower_bound <- q1 - 1.5 * iqr
  upper_bound <- q3 + 1.5 * iqr
  outliers <- sum(x < lower_bound | x > upper_bound, na.rm = TRUE)  # Count of outliers

  c(
    n = sum(!is.na(x)),  # Number of observations
    mean = mean(x, na.rm = TRUE),
    sd = sd(x, na.rm = TRUE),
    median = median(x, na.rm = TRUE),
    min = min(x, na.rm = TRUE),
    max = max(x, na.rm = TRUE),
    missing = sum(is.na(x)),  # Count of missing values
    outliers = outliers  # Count of outliers
  )
})

# Transpose for readability and convert to data frame
summary_stats <- as.data.frame(t(summary_stats))

# Add descriptive column names
colnames(summary_stats) <- c("n", "Mean", "SD", "Median", "Min", "Max", "Missing", "Outliers")

# Print the summary statistics in a table format
knitr::kable(summary_stats, caption = "Table 1: Summary Statistics of Key Variables")
```

Table 1: Table 1: Summary Statistics of Key Variables

	n	Mean	SD	Median	Min	Max	Missing	Outliers
MentHlth	70692	3.7520370	8.1556266	0.0	0	30	0	11816
PhysHlth	70692	5.8104170	10.0622605	0.0	0	30	0	10624
Smoker	70692	0.4752730	0.4993917	0.0	0	1	0	0
PhysActivity	70692	0.7030357	0.4569239	1.0	0	1	0	0
Diabetes_binary	70692	0.5000000	0.5000035	0.5	0	1	0	0

3. Hypothesis Testing Continuous Variables (t-tests)

```r
# Compare MentHlth and PhysHlth between diabetic and non-diabetic individuals

# If the average number of poor mental health differs significantly between those with and without diab
t_test_menthlth <- t.test(MentHlth ~ Diabetes_binary, data = diabetes_data)
t_test_menthlth
```

```
##
##  Welch Two Sample t-test
##
## data:  MentHlth by Diabetes_binary
## t = -23.227, df = 67626, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -1.539325 -1.299751
## sample estimates:
## mean in group 0 mean in group 1
##        3.042268        4.461806
```

```r
# if the average number of poor physical health differs significantly between those with and without di
t_test_physhlth <- t.test(PhysHlth ~ Diabetes_binary, data = diabetes_data)
t_test_physhlth
```

```
##
##  Welch Two Sample t-test
##
## data:  PhysHlth by Diabetes_binary
## t = -57.985, df = 64069, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -4.433070 -4.143176
## sample estimates:
## mean in group 0 mean in group 1
##        3.666355        7.954479
```

The p-values are extremely small.

There is a statistically significant difference in mental health between individuals with and without diabetes. On average, individuals with diabetes report more poor mental health days.

There is a statistically significant difference in physical health between individuals with and without diabetes. On average, individuals with diabetes report significantly more poor physical health days.

```r
# Summary: means for Mental Health and Physical Health by Diabetes status
summary_stats <- diabetes_data %>%
  group_by(Diabetes_binary) %>%
  summarise(
    mean_menthlth = mean(MentHlth, na.rm = TRUE),
    mean_physhlth = mean(PhysHlth, na.rm = TRUE)
  )
summary_stats
```

```
## # A tibble: 2 x 3
##   Diabetes_binary mean_menthlth mean_physhlth
##             <dbl>         <dbl>         <dbl>
## 1               0          3.04          3.67
## 2               1          4.46          7.95
```

Categorical Variables (Chi-square Tests)

```r
# Test association between lifestyle (smoking and physical activity) and diabetes
chisq_smoker <- chisq.test(table(diabetes_data$Smoker, diabetes_data$Diabetes_binary))
chisq_smoker
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(diabetes_data$Smoker, diabetes_data$Diabetes_binary)
## X-squared = 522.48, df = 1, p-value < 2.2e-16
```

```r
chisq_physactivity <- chisq.test(table(diabetes_data$PhysActivity, diabetes_data$Diabetes_binary))
chisq_physactivity
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(diabetes_data$PhysActivity, diabetes_data$Diabetes_binary)
## X-squared = 1779, df = 1, p-value < 2.2e-16
```

The p-values are extremely small. Both lifestyle factors (smoking and physical activity) are significantly associated with diabetes status.

Summary of all categorical variables (Smoker, PhysActivity, Diabetes_binary)

```r
summarize_categorical <- function(data, var) {
  data %>%
    group_by(!!sym(var)) %>%
    summarise(
      N = n(),
      Percentage = round((N / nrow(data)) * 100, 2)
    )
}

# Summarize for each categorical variable
```

```r
summary_smoker <- summarize_categorical(diabetes_data, "Smoker")
summary_physactivity <- summarize_categorical(diabetes_data, "PhysActivity")
summary_diabetes <- summarize_categorical(diabetes_data, "Diabetes_binary")

summary_smoker
```

```
## # A tibble: 2 x 3
##   Smoker     N Percentage
##    <dbl> <int>      <dbl>
## 1      0 37094       52.5
## 2      1 33598       47.5
```

```r
summary_physactivity
```

```
## # A tibble: 2 x 3
##   PhysActivity     N Percentage
##          <dbl> <int>      <dbl>
## 1            0 20993       29.7
## 2            1 49699       70.3
```

```r
summary_diabetes
```

```
## # A tibble: 2 x 3
##   Diabetes_binary     N Percentage
##             <dbl> <int>      <dbl>
## 1               0 35346         50
## 2               1 35346         50
```

The percentage of Diabetes vs Non-diabetes in the dataset is 50% vs 50%

```r
# Simplify dataset by taking the mean of continuous variables
simplified_data <- diabetes_data %>%
  group_by(Diabetes_binary) %>%
  summarise(
    mean_menthlth = mean(MentHlth, na.rm = TRUE),
    mean_physhlth = mean(PhysHlth, na.rm = TRUE),
    Smoker = mean(Smoker, na.rm = TRUE),
    PhysActivity = mean(PhysActivity, na.rm = TRUE)
  )

# Use knitr::kable to display the summarized data as a table
knitr::kable(
  simplified_data,
  caption = "Table 2: Simplified Summary Data Grouped by Diabetes Status"
)
```

Table 2: Table 2: Simplified Summary Data Grouped by Diabetes Status

Diabetes_binary	mean_menthlth	mean_physhlth	Smoker	PhysActivity
0	3.042268	3.666355	0.4323261	0.7755333
1	4.461806	7.954479	0.5182199	0.6305381

```r
# Correlation between MentHlth and PhysHlth
correlation <- cor.test(diabetes_data$MentHlth, diabetes_data$PhysHlth)
correlation
```

```
##
##  Pearson's product-moment correlation
##
## data:  diabetes_data$MentHlth and diabetes_data$PhysHlth
## t = 109.32, df = 70690, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3739483 0.3865597
## sample estimates:
##       cor
## 0.3802717
```

There is a statistically significant (small p-value), moderate positive correlation (correlation coefficient > 0) between poor mental health and poor physical health.

```r
# Split dataset into training (50%) and testing (50%)
set.seed(42)  # For reproducibility
train_indices <- sample(1:nrow(diabetes_data), nrow(diabetes_data) / 2)
train_data <- diabetes_data[train_indices, ]
test_data <- diabetes_data[-train_indices, ]

# Fit logistic regression on training data
logistic_model <- glm(Diabetes_binary ~ MentHlth + PhysHlth + Smoker + PhysActivity,
                      data = train_data, family = binomial)
summary(logistic_model)
```

```
##
## Call:
## glm(formula = Diabetes_binary ~ MentHlth + PhysHlth + Smoker +
##     PhysActivity, family = binomial, data = train_data)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.0544186  0.0253016   2.151   0.0315 *
## MentHlth     -0.0004042  0.0014994  -0.270   0.7875
## PhysHlth      0.0401612  0.0012919  31.086   <2e-16 ***
## Smoker        0.2186310  0.0221486   9.871   <2e-16 ***
## PhysActivity -0.5238769  0.0248477 -21.084   <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: 49000  on 35345  degrees of freedom
## Residual deviance: 46727  on 35341  degrees of freedom
## AIC: 46737
##
## Number of Fisher Scoring iterations: 4
```

```r
model_summary <- summary(logistic_model)

# Extract coefficients
coefficients <- as.data.frame(model_summary$coefficients)

# Rename columns for clarity
colnames(coefficients) <- c("Estimate", "Std. Error", "z-value", "p-value")

# Print table using knitr::kable
knitr::kable(
  coefficients,
  caption = "Table 3: Logistic Regression Results"
)
```

Table 3: Table 3: Logistic Regression Results

	Estimate	Std. Error	z-value	p-value
(Intercept)	0.0544186	0.0253016	2.1507943	0.0314924
MentHlth	-0.0004042	0.0014994	-0.2695663	0.7874939
PhysHlth	0.0401612	0.0012919	31.0862345	0.0000000
Smoker	0.2186310	0.0221486	9.8711169	0.0000000
PhysActivity	-0.5238769	0.0248477	-21.0835040	0.0000000

For mental health, the estimate is pretty small, meaning that the log-odds of diabetes decrease very slightly for each additional day of poor mental health. And the p-value = 0.7875, which is not significant. This shows that mental health does not significantly contribute to predicting diabetes when adjusted for other factors.

For physical health, the estimate is 0.04, meaning that each additional day of poor physical health increases the log-odds of diabetes by 0.04. The Odds Ratio is 1.041, which means each additional day of poor physical health increases the odds of diabetes by ~4.1%. P-value is extremely small, showing the relationship is statistically significant.

For smoking, being a smoker increases the log-odds of diabetes by 0.219 (Odds Ratio: 1.244), meaning that smokers have 24.4% higher odds of having diabetes compared to non-smokers. The small P-value shows that smoking is strongly associated with diabetes.

For physical activity, being physically active decreases the log-odds of diabetes by 0.524 (Odds Ratio: 0.592), representing that physically active individuals have 59.2% lower odds of having diabetes. The extremely small P-value shows that physical activity is a significantly strong predictive factor of diabetes.

```r
# Predict on the test set
test_data$predicted_prob <- predict(logistic_model, newdata = test_data, type = "response")
test_data$predicted_class <- ifelse(test_data$predicted_prob > 0.5, 1, 0)
```

```r
# Confusion matrix
confusion_matrix <- table(test_data$Diabetes_binary, test_data$predicted_class)
confusion_matrix
```

```
##
##         0     1
##   0 12335  5388
##   1  8540  9083
```

```r
# Accuracy
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
accuracy
```

```
## [1] 0.6059526
```

A moderate level of accuracy. The dataset is balanced (50% diabetic, 50% non-diabetic), so random guessing would yield an accuracy of around 50%. The model's accuracy of 60.6% is better than that, indicating some predictive power.

```r
# Sensitivity (Recall): ability correctly identify people with a disease (true positives).

sensitivity <- confusion_matrix[2, 2] / sum(confusion_matrix[2, ])
sensitivity
```

```
## [1] 0.515406
```

The model correctly identifies 51.5% of individuals with diabetes but misses nearly half of them. To Improve Sensitivity: 1) Adjust the classification threshold (default is 0.5) to favor capturing more positive cases. 2) Add more features or interactions to better capture the variability in diabetic cases.

```r
# Specificity: ability to correctly identify people without a disease (true negatives)
specificity <- confusion_matrix[1, 1] / sum(confusion_matrix[1, ])
specificity
```

```
## [1] 0.6959883
```

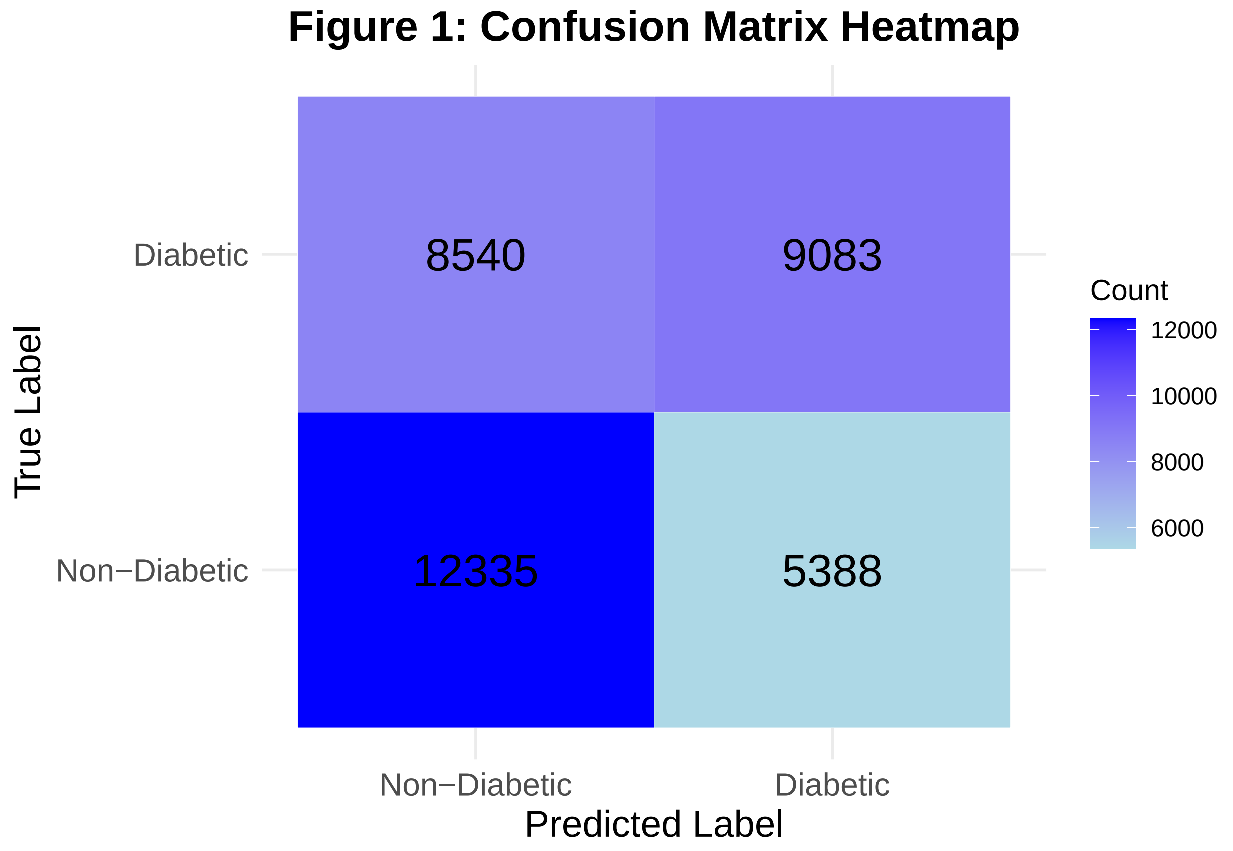
The model correctly identifies ~70% of non-diabetic individuals but misclassifies ~30% of them as diabetic. The model performs better in identifying non-diabetic cases (higher specificity)

```r
library(ggplot2)

# Define the confusion matrix
confusion_matrix <- matrix(c(12335, 5388, 8540, 9083), nrow = 2, byrow = TRUE)
rownames(confusion_matrix) <- c("Non-Diabetic", "Diabetic")  # True labels
colnames(confusion_matrix) <- c("Non-Diabetic", "Diabetic")  # Predicted labels

# Convert confusion matrix to data frame for ggplot
confusion_df <- as.data.frame(as.table(confusion_matrix))
colnames(confusion_df) <- c("True_Label", "Predicted_Label", "Count")
```

```r
# Plot heatmap
ggplot(confusion_df, aes(x = Predicted_Label, y = True_Label, fill = Count)) +
  geom_tile(color = "white") +  # Create heatmap tiles
  geom_text(aes(label = Count), color = "black", size = 6) +  # Add text labels for counts
  scale_fill_gradient(low = "lightblue", high = "blue") +  # Color gradient
  labs(
    title = "Figure 1: Confusion Matrix Heatmap",
    x = "Predicted Label",
    y = "True Label"
  ) +
  theme_minimal() +  # Minimal theme
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    axis.text = element_text(size = 12),
    axis.title = element_text(size = 14)
  )
```



6. Metrics (AUC and Confidence Intervals)

```r
# Evaluate model performance using AUC
library(pROC)
roc_curve <- roc(test_data$Diabetes_binary, test_data$predicted_prob)
```

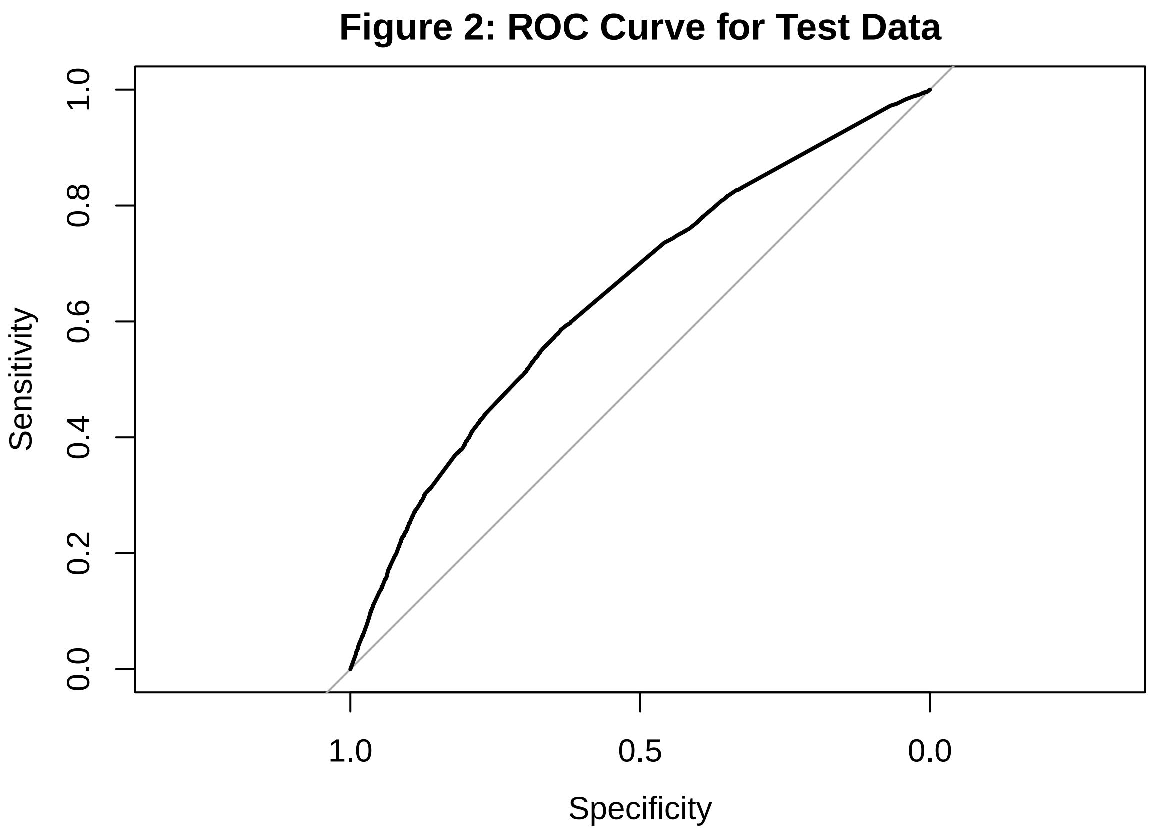
```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
auc_value <- auc(roc_curve)

auc_value
```

```
## Area under the curve: 0.6468
```

```
# Plot the ROC curve
plot(roc_curve, main = "Figure 2: ROC Curve for Test Data")
```



The AUC of 0.6468 indicates the model has fair predictive power for identifying diabetes.