STA 478 HW 2

Jared Brotamonte

9/13/2023

Exercise 1

(a)

```
setwd("C:/Users/jkbro/OneDrive/Desktop/STA 478/Homework/HW 2")
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.1.3
```

```
## Loading required package: ggplot2
```

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

```
## Warning: package 'lattice' was built under R version 4.1.3
```

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.1.3
```

```
# read in data
wine_data <- read.csv("C:/Users/jkbro/OneDrive/Desktop/STA 478/Homework/HW 2/wine.csv")
head(wine_data)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.0             0.27        0.36           20.7     0.045
## 2           6.3             0.30        0.34            1.6     0.049
## 3           8.1             0.28        0.40            6.9     0.050
## 4           7.2             0.23        0.32            8.5     0.058
## 5           7.2             0.23        0.32            8.5     0.058
## 6           8.1             0.28        0.40            6.9     0.050
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  45                  170  1.0010 3.00      0.45     8.8
## 2                  14                  132  0.9940 3.30      0.49     9.5
## 3                  30                   97  0.9951 3.26      0.44    10.1
## 4                  47                  186  0.9956 3.19      0.40     9.9
## 5                  47                  186  0.9956 3.19      0.40     9.9
## 6                  30                   97  0.9951 3.26      0.44    10.1
##   quality
## 1       6
## 2       6
## 3       6
## 4       6
## 5       6
## 6       6
```

```
# set aside 17th row to estimate
wine_to_predict <- wine_data[17, ]

# remove 17th row
wine_data <- wine_data[-17, ]

# perform knn with k=3
k_val <- 3
knn_model <- knn3(quality ~., data = wine_data, k = k_val)

# predict quality
predicted <- predict(knn_model, wine_to_predict)
predicted
```

```
##      3         4         5         6 7 8 9
## [1,] 0 0.3333333 0.3333333 0.3333333 0 0 0
```

(b)

```
# do same for k-15
k_val <- 15
knn_model <- knn3(quality ~ ., data = wine_data, k = k_val)

# predict quality
predicted <- predict(knn_model, wine_to_predict)
predicted
```

```
##      3          4   5   6         7 8 9
## [1,] 0 0.06666667 0.2 0.6 0.1333333 0 0
```

(c)

```
# Find actual quality
actual <- wine_to_predict$quality
actual
```

```
## [1] 6
```

Exercise 2

```
# load libraries
library(ISLR2)
```

```
## Warning: package 'ISLR2' was built under R version 4.1.3
```

```
data("Default")
```

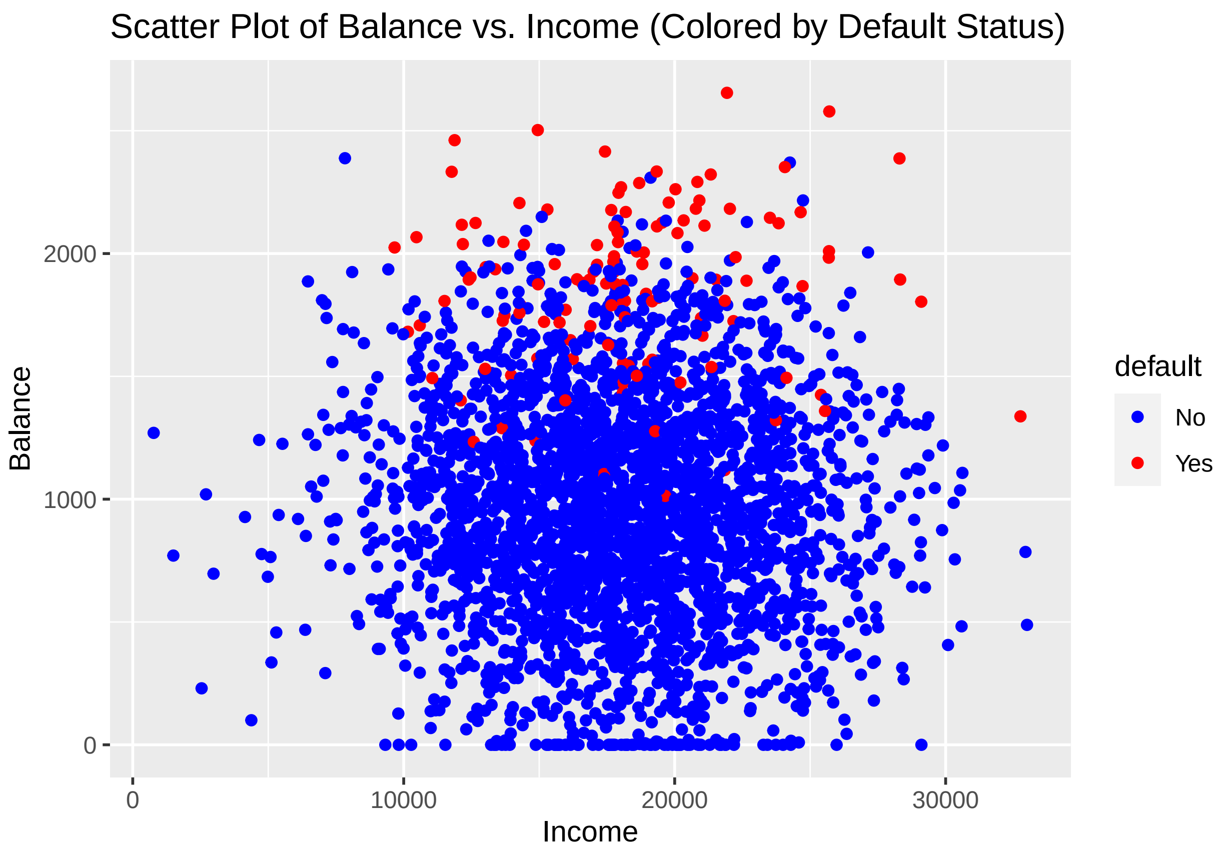
(a)

```
# filter to keep only students
students_data <- Default[Default$student == "Yes",]
head(students_data)
```

```
##    default student   balance    income
## 2       No     Yes  817.1804 12106.135
## 6       No     Yes  919.5885  7491.559
## 8       No     Yes  808.6675 17600.451
## 11      No     Yes    0.0000 21871.073
## 12      No     Yes 1220.5838 13268.562
## 18      No     Yes  527.5402 17636.540
```

(b)

```
library(ggplot2)
ggplot(data = students_data, aes(x = income, y = balance, color = default)) +
  geom_point() +
  scale_color_manual(values = c("No" = "blue", "Yes" = "red")) +
  labs(x = "Income", y = "Balance") +
  ggtitle("Scatter Plot of Balance vs. Income (Colored by Default Status)")
```



It's hard to deduce any information from this plot because it is so cluttered.

(c)

```
# Define the test instance
test_instance <- data.frame(income = 18000, balance = 1900)

# perform knn with k=11
k_val <- 11
knn_model <- knn3(default ~ income + balance, data = students_data, k=k_val)
knn_model
```

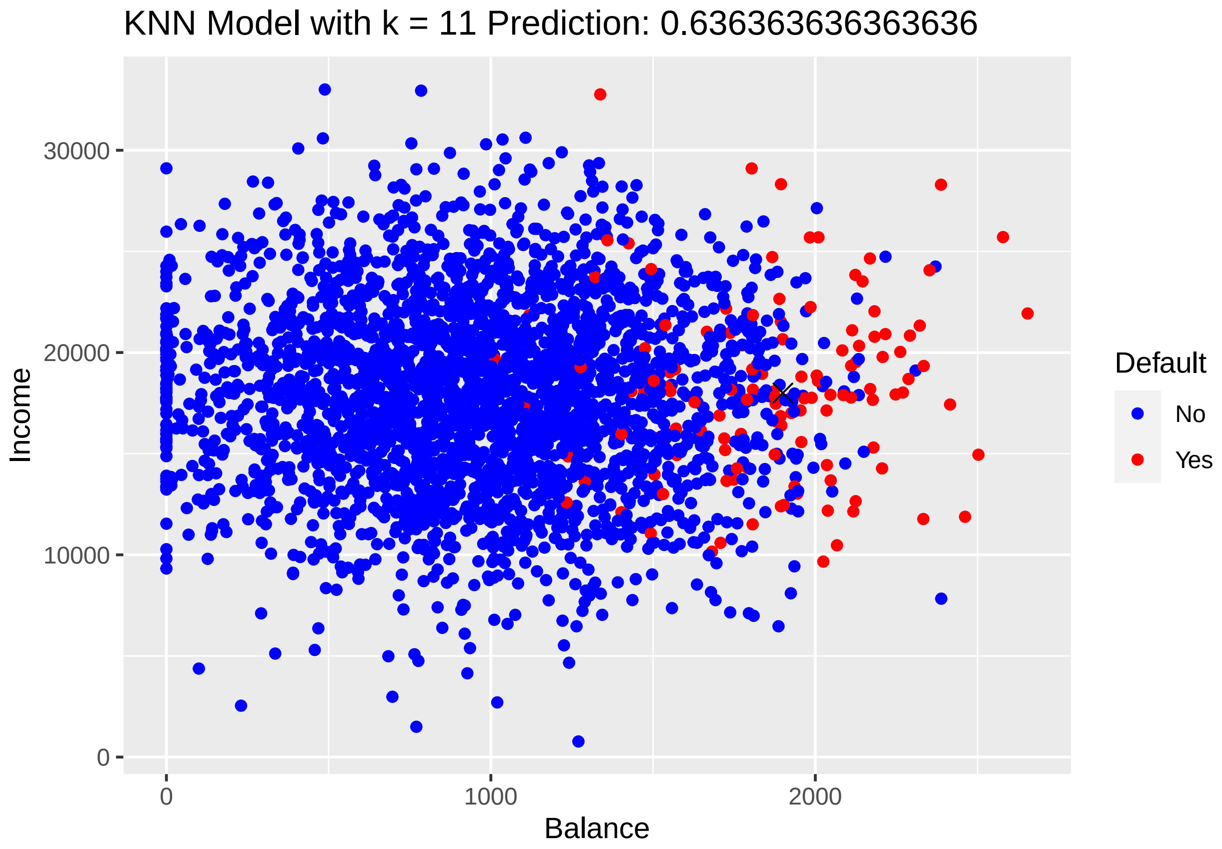
```
## 11-nearest neighbor model
## Training set outcome distribution:
##
##   No  Yes
## 2817  127
```

```
predicted <- predict(knn_model,newdata = test_instance)

# Create a dataframe for plotting
plot_data <- data.frame(income = students_data$income, balance = students_data$balance, Default = studen

# Create a scatter plot
```

```
ggplot(data = plot_data, aes(x = balance, y = income, color = Default)) +
  geom_point() +
  geom_point(data = test_instance, aes(x = balance, y = income), color = "black", size = 3, shape = 4) +
  scale_color_manual(values = c("No" = "blue", "Yes" = "red")) +
  labs(x = "Balance", y = "Income") +
  ggtitle(paste("KNN Model with k =", k_val, "Prediction:", predicted))
```



```
# perform knn with k=55
k_val <- 55
knn_model <- knn3(default ~ income + balance, data = students_data, k=k_val)
knn_model
```

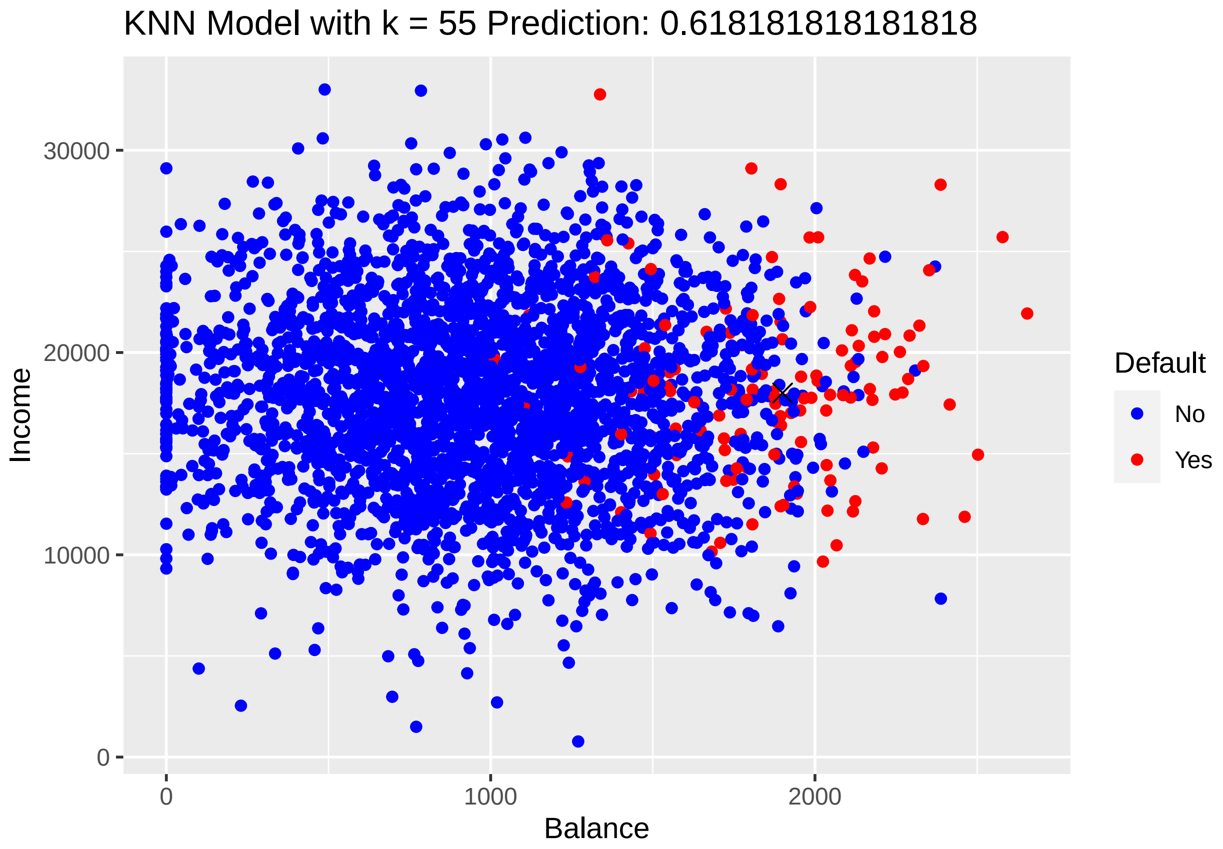
```
## 55-nearest neighbor model
## Training set outcome distribution:
##
##   No  Yes
## 2817  127
```

```
predicted <- predict(knn_model,newdata = test_instance)

# Create a dataframe for plotting
plot_data <- data.frame(income = students_data$income, balance = students_data$balance, Default = stude

# Create a scatter plot
```

```
ggplot(data = plot_data, aes(x = balance, y = income, color = Default)) +
  geom_point() +
  geom_point(data = test_instance, aes(x = balance, y = income), color = "black", size = 3, shape = 4) +
  scale_color_manual(values = c("No" = "blue", "Yes" = "red")) +
  labs(x = "Balance", y = "Income") +
  ggtitle(paste("KNN Model with k =", k_val, "Prediction:", predicted))
```

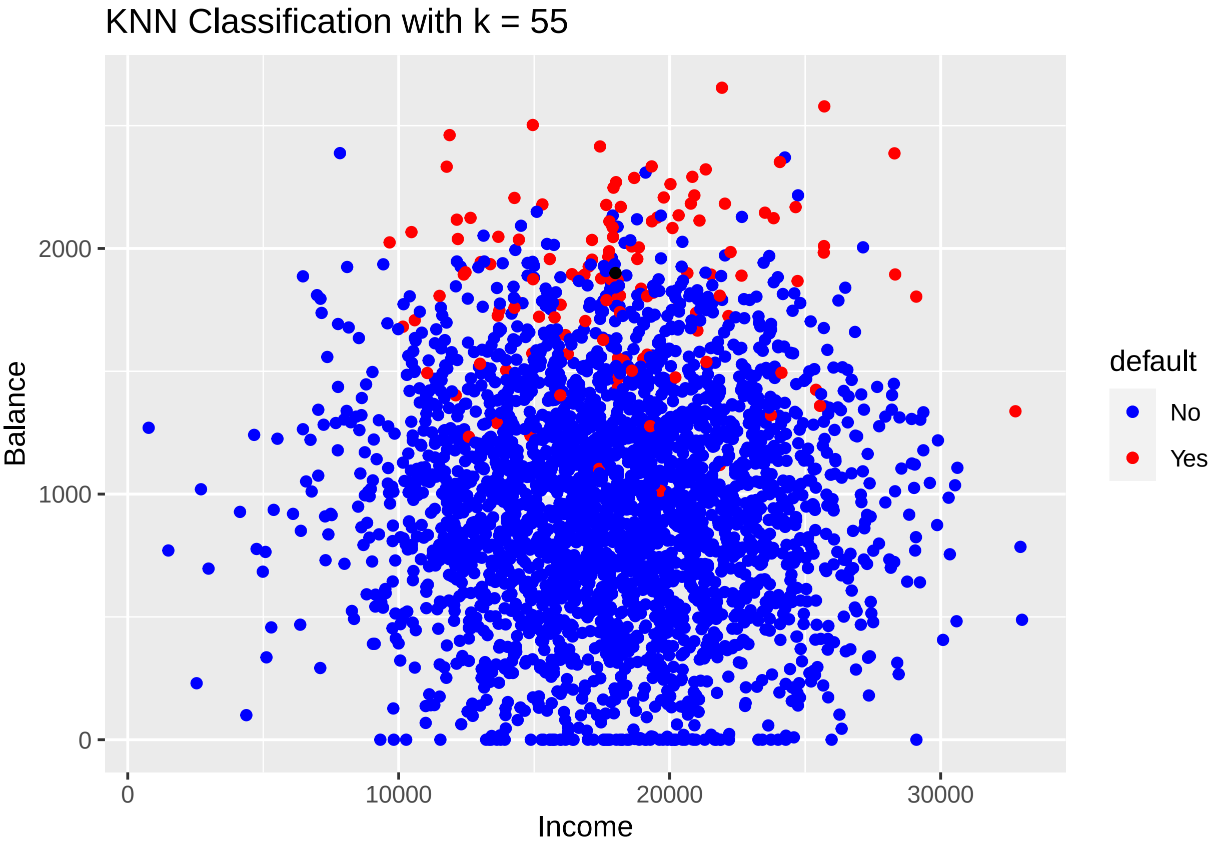


(d)

```
# repeate with k-55
k_val <- 55
knn_model <- knn3(default ~ income + balance, data = students_data, k=k_val)
knn_model
```

```
## 55-nearest neighbor model
## Training set outcome distribution:
##
##   No  Yes
## 2817  127
```

```
ggplot() +
  geom_point(data = students_data, aes(x = income, y = balance, color = default)) +
  geom_point(data = test_instance, aes(x = income, y = balance, color = knn_model$class)) +
  scale_color_manual(values = c("No" = "blue", "Yes" = "red")) +
  labs(x = "Income", y = "Balance") +
  ggtitle("KNN Classification with k = 55")
```



(e)

The k-55 should be the more flexible model thus having a variance while the k-11 would be less flexible and have a higher bias.