Business Analytics Final Project Task 2

MGT 7104

Yacine Ndiaye

May 5th,2025

Instructor: Jaime Velasco Sanchez

Richmond The American University in London

Word Count: 1115

TABLE OF CONTENTS

TASK 2

INTRODUCTION

In an effort to boost customer participation in e-banking as part of the rollout for its online platform, Sandhills bank has decided to offer a higher interest rate on savings account to customers who opt in for comprehensive paperless services. To evaluate the potential success rate of this promotion, Sandhills seeks to estimate how many of its 200 current customers who are not yet enrolled for e-banking would likely accept the offer.

To make this prediction, data from a similar promotion executed by a partner institution named "plains bank" has been provided. This dataset contains information on 1000 customers, each with their average monthly checking account balance, age and e-banking enrolment status.

This paper will be using Sandhill's CSV file which contains information about its 200 customers for whom the promotion has yet to be launched. This dataset only covers information regarding their average monthly checking account balance and the customer's age; the sandhills dataset is missing the "enrol" column. To counter that issue, synthetic data will be generated.

The synthetic "plain csv file" is based off the "sandhills csv file" and includes 1000 rows with three variables: balance (random between 100-1000), age (random between 18-80) and enrol (target variable). After this procedure, the KNN model will be used on a 100% of the data in the plain's csv file for training and validation; none of the data will be used as a test set which will lead to finding the best "K" that maximizes AUC. The model will then be applied to the 200 sandhills bank customers from the Sandhill dataset to predict enrolment. Finally, a count of the predicted enrolment using a 0.5 probability cutoff will be provided.

The KNN algorithm is a non-parametric machine learning algorithm mainly used for regression and classification problems. It bases its classification and predictions off the proximity of similar items by finding the nearest neighbour to a particular point and predicting class of value based on the characteristics of the neighbours of that particular point. (LaViale, 2023)

The packages used to conduct this procedure are: "class" for classification; "pROC" for a ROC analysis; "caret" for cross validation; "ggplot2" for visualization and "dplyr" for data manipulation.

All procedures used in the analysis of the data for this case will be listed and explained.

TASK 2.1

The synthetically generated “plains csv” file was used to determine the value of “K” that maximizes the AUC in a validation procedure.

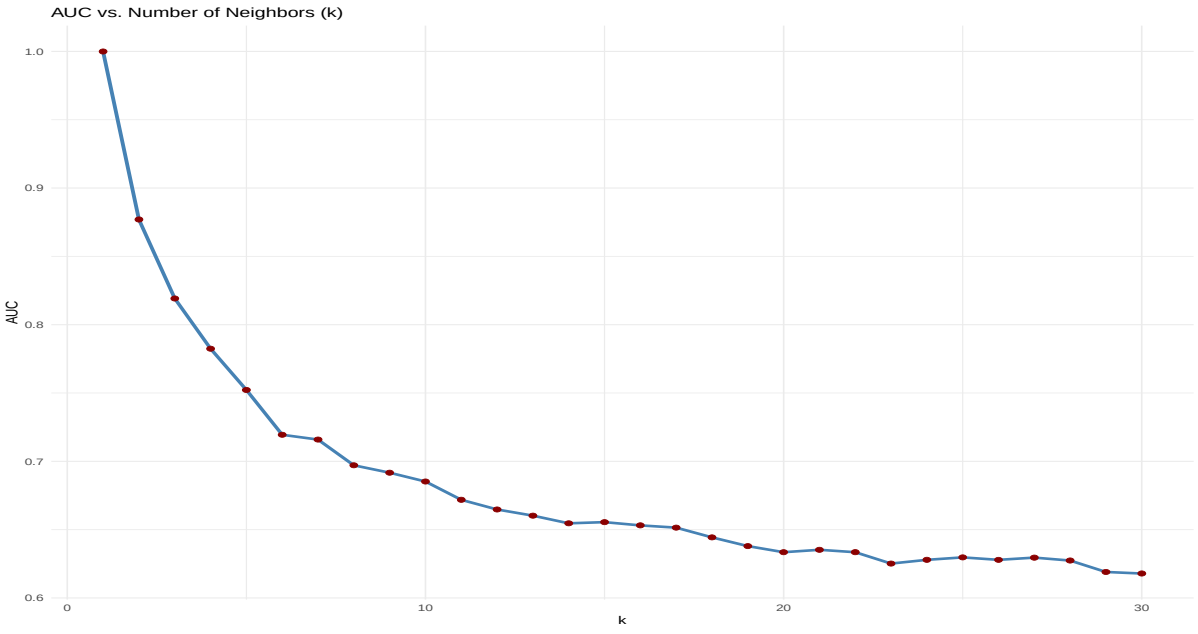
After generating the dataset, its content was normalized to put all features on the same scale {0,1}; As a distance-based algorithm, KNN requires normalization; it’s a data preprocessing technique used to rescale numerical features to a standard range, ensuring balance and stability. (Abdalla, Altaf, 2023)

The values of “K” were tried from 1 to 30 after which the “KNN” algorithm is applied using all the normalized data from the plains dataset to predict the target variable “enrol”. This is done for different values of “K”. The model predicts the class of each data point based on its K nearest neighbours and calculates the predicted probability for the class 1 (enrolled).

The AUC is a measure of the model’s performance that evaluates how well the model discriminates against classes (Dash, 2022). The peak AUC score represents the best K value which in this case is 1.0, it will be used for the final KNN model.

The graph visualization below shows how the AUC score changes as the number of neighbours (K) varies. The X axis represents “K” and shows the number of neighbours considered in the KNN algorithm (1-30). The Y axis is representative of the AUC which reflects the model’s performance. It ranges from 0 to 1 where 1 is a perfect classification, 0.5 is random guessing and <0.5 means that the model’s performance is mediocre.

When K is between 30-20, AUC is considerably low barely above 0.6. when K is between 20-10, AUC starts to steadily pick up. when K is between 10-0 AUC dramatically takes off and reaches its peak at 1.0 where K is closer to 0.

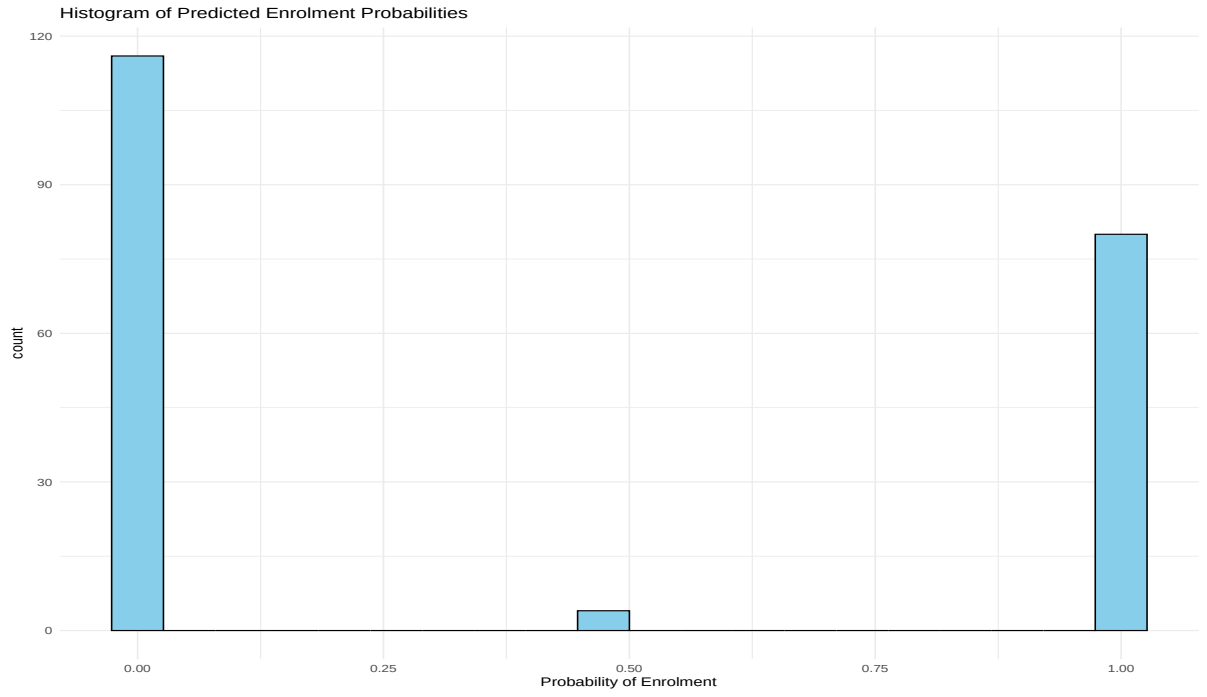


TASK 2.2

Prior to the application of the final model selected in the previous task to the 200 observations in the sandhills file, the data from the latter has been normalized using the same scale as the plains data, to ensure consistency. After which, the final KNN model was applied to predict the "enrol" variable for the sandhills data using the best K (1).

In order for the final model to make predictions on how many of sandhill bank's 200 customers will enrol for paperless banking using a cutoff value of 0.5, the model uses the normalized "balance" and "age" features from both the plains data as training data and sandhill data as test data; it then returns predicted labels and the probabilities for class 1 (enrolled) are extracted. The final predicted enrolment is computed by classifying customers whose probability is greater than 0.5 as 1 (enrolled).

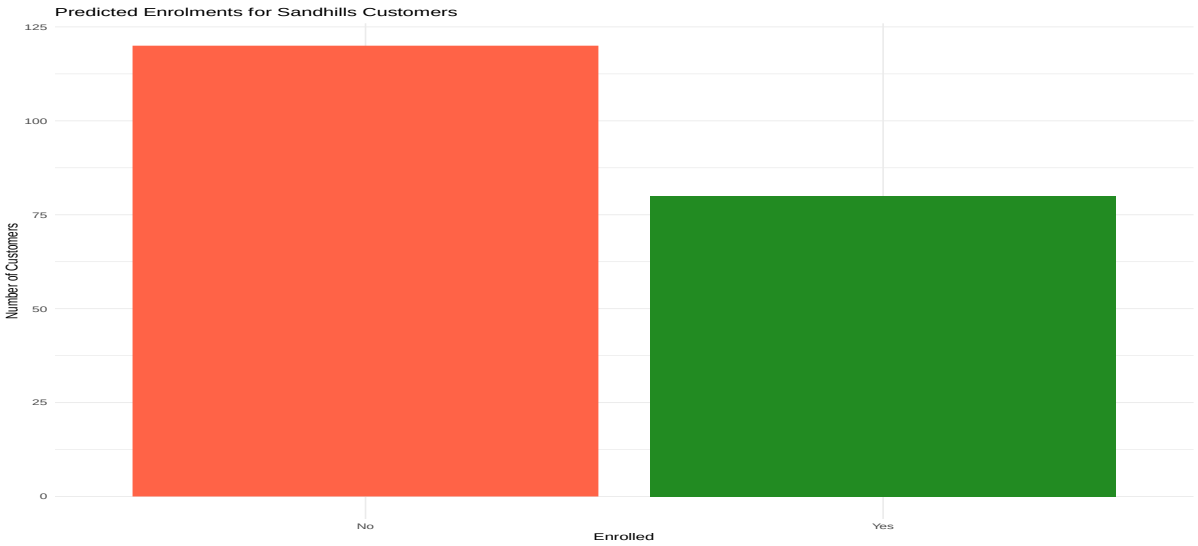
This can be seen in the histogram below. It's a visualization of the distribution of predicted probabilities for enrolment. The graph shows that the model predicts the non-enrolment of majority of customers (100+) and the enrolment of less than 90 customers. The middle range counts very few customers meaning that the model is confident in its predictions.



TASK 2.3

The chart below helps visualize the number of predicted enrolled and non-enrolled customers in the sandhills dataset. According to this graph, the amount of non-enrolled customers is well above 100 and considerably exceed the number of enrolled customers who are barely above 75.

The exact number of predicted enrolments with a probability threshold of 0.5 made by the model is 80. The number is quite low as it doesn't account for half of the number of customers.



CONCLUSION

This analysis aimed to predict the enrolment rate of Sandhills bank's 200 customers into paperless banking services using a KNN model which was trained on synthetic data from a partner institution named plains bank then applied to sandhills' customer data. The KNN model achieved a perfect AUC score of 1.0, indicating optimal performance during validation. Suggesting that the model perfectly classified enrolment status in the synthetic dataset. However, such a high AUC might Also hint at potential overfitting as real world data rarely allows for perfect separation. Further validation with real data would be most beneficial. The model's prediction of the enrolment of 40% of customers (80 out of 200) using a probability threshold of 0.5 aligns with the assumed enrolment distribution in the synthetic data (40%). These results suggests that Sandhills bank could expect moderate uptake of its paperless banking promotion however the bank should consider launching a small-scale pilot to validate predictions as well as incorporating additional variables such as transaction frequency to improve accuracy.

REFERENCES

Abdalla, H.I., Altaf, A. 2023. "The Impact of Data Normalization on KNN Rendering. In: Hassanien, A., Rizk, R.Y., Pamucar, D., Darwish, A., Chang, KC. (eds)". *Springer nature link*. AISI 2023. Lecture Notes on Data Engineering and Communications Technologies, vol 184. Available at: https://link.springer.com/chapter/10.1007/978-3-031-43247-7_16#citeas Accessed: 03/05/2025

Dash, S., 2022. "Understanding the ROC and AUC intuitively". *Medium*. Available at: https://medium.com/@shaileydash/understanding-the-roc-and-auc-intuitively-31ca96445c02 Accessed: 03/05/2025

LaViale, 2023. "Deep Dive on KNN: Understanding and Implementing the K-Nearest Neighbours Algorithm". *Arize*. Available at: https://arize.com/blog-course/knn-algorithm-k-nearest-neighbor/ Accessed: 03/05/2025

INDEX

KNN: K Nearest Neighbour

AUC: Area Under the Curve

ROC: Receiver Operating Characteristics

R-CODE TASK 2

```
# Load necessary libraries
install.packages(c("class", "pROC", "caret", "ggplot2", "dplyr"))
library(class)
library(pROC)
library(caret)
library(ggplot2)
library(dplyr)

# Load Sandhills customer data
sandhills <- read.csv("SH.csv")



# Step 1: Generate synthetic Plains Bank training data
set.seed(123)
n_train <- 1000
plains <- data.frame(
  Balance = round(runif(n_train, 100, 1000)),
  Age = round(runif(n_train, 18, 80)),
  Enrol = sample(c(0, 1), n_train, replace = TRUE, prob = c(0.6, 0.4)) # Assume 40%
enrolment
)



# Step 2: Normalize features
normalize <- function(x) (x - min(x)) / (max(x) - min(x))
plains_norm <- plains
plains_norm$Balance <- normalize(plains$Balance)
plains_norm$Age <- normalize(plains$Age)
```

```
# Step 3: Cross-validation to find best k (now done with the whole dataset without a test set)
set.seed(123)
k_values <- 1:30
auc_scores <- numeric(length(k_values))

for (i in seq_along(k_values)) {
  aucs <- c()

  # Training and testing are done using all data in the current fold
  train_labels <- plains$Enrol
  pred <- knn(plains_norm[, c("Balance", "Age")], plains_norm[, c("Balance", "Age")],
train_labels, k = k_values[i], prob = TRUE)
  probs <- ifelse(pred == "1", attr(pred, "prob"), 1 - attr(pred, "prob"))

  aucs <- c(aucs, auc(roc(train_labels, probs)))

  auc_scores[i] <- mean(aucs)
}




# Step 4: Plot AUC vs. k
auc_df <- data.frame(k = k_values, AUC = auc_scores)
ggplot(auc_df, aes(x = k, y = AUC)) +
  geom_line(color = "steelblue", linewidth = 1.2) +
  geom_point(color = "darkred", linewidth = 2) +
  ggtitle("AUC vs. Number of Neighbors (k)") +
  theme_minimal()




# Step 5: Select best k
best_k <- k_values[which.max(auc_scores)]
```

```
cat("Best k based on AUC:", best_k, "\n")

# Step 6: Normalize Sandhills data using same scale
sandhills_norm <- sandhills
sandhills_norm$Balance <- normalize(c(plains$Balance,
sandhills$Balance))[(n_train+1):(n_train+200)]
sandhills_norm$Age <- normalize(c(plains$Age, sandhills$Age))[(n_train+1):(n_train+200)]

# Step 7: Apply final KNN model to Sandhills customers
final_pred <- knn(plains_norm[, c("Balance", "Age")], sandhills_norm[, c("Balance",
"Age")],
                  plains$Enrol, k = best_k, prob = TRUE)

final_probs <- ifelse(final_pred == "1", attr(final_pred, "prob"), 1 - attr(final_pred, "prob"))
predicted_enrol <- ifelse(final_probs > 0.5, 1, 0)

# Step 8: Visualizations
# (a) Histogram of predicted probabilities
ggplot(data.frame(Prob = final_probs), aes(x = Prob)) +
  geom_histogram(bins = 20, fill = "skyblue", color = "black") +
  ggtitle("Histogram of Predicted Enrolment Probabilities") +
  xlab("Probability of Enrolment") +
  theme_minimal()

# (b) Bar chart of final predictions
enrol_df <- data.frame(Enrolled = factor(predicted_enrol, levels = c(0, 1), labels = c("No",
"Yes")))
ggplot(enrol_df, aes(x = Enrolled)) +
  geom_bar(fill = c("tomato", "forestgreen")) +
  ggtitle("Predicted Enrolments for Sandhills Customers") +
  ylab("Number of Customers") +
  theme_minimal()

# Step 9: Count predicted enrolments
```

```
cat("Number of predicted enrolments (prob > 0.5):", sum(predicted_enrol), "\n")
```