# loading libraries

library(caret)

library(mlbench)

library(magrittr)

library(dplyr)

library(ggplot2)

library(ggpubr)

# setting ggplot Themes

theme\_set(theme\_pubr())

# load the CSV file from the local directory

dataset <- read.csv("C:/Users/HOME/Downloads/College\_admission.csv", header=TRUE)

# Check first few rows

head(dataset)

# Summary the data

summary(dataset)

# summarize the admission distribution

percentage <- prop.table(table(dataset$admit)) \* 100

cbind(freq=table(dataset$admit), percentage=percentage)

#Find the missing values. (if any, perform missing value treatment)

is.na(dataset)

sum(is.na(dt))

# No missing values in dataset

#Find outliers (if any, then perform outlier treatment)

outliers <- boxplot(dataset$disp, plot=FALSE)$out

dataset[which(dataset$disp %in% outliers),]

# No outliers in dataset

#Find the structure of the data set

str(dataset)

# list types for each attribute

sapply(dataset, class)

#transform the numeric data type to factor and vice-versa.

dataset$admit=as.factor(dataset$admit)

# list types after tansfer to factor

sapply(dataset, class)

#Find whether the data is normally distributed or not. Use the plot to determine the same.

x <- dataset[,2:7]

y <- dataset[,1]

# density plots for each attribute by class value

scales <- list(x=list(relation="free"), y=list(relation="free"))

featurePlot(x=x, y=y, plot="density", scales=scales)

# create a list of 80% of the rows in the original dataset we can use for training

validation\_index <- createDataPartition(dataset$admit, p=0.80, list=FALSE)

# select 20% of the data for validation

validation <- dataset[-validation\_index,]

# use the remaining 80% of data to training and testing the models

dataset <- dataset[validation\_index,]

# Use variable reduction techniques to identify significant variables.

set.seed(7)

# calculate correlation matrix

control <- trainControl(method="repeatedcv", number=10, repeats=3)

# train the model

model <- train(admit~., data=dataset, method="lvq", preProcess="scale", trControl=control)

# estimate variable importance

importance <- varImp(model, scale=FALSE)

# summarize importance

print(importance)

# plot importance

plot(importance)

#Run logistic model to determine the factors that influence the admission process of a student (Drop insignificant variables)

#Calculate the accuracy of the model and run validation techniques.

#Try other modelling techniques like decision tree and SVM and select a champion model

#Determine the accuracy rates for each kind of model

#Select the most accurate model

#Identify other Machine learning or statistical techniques

set.seed(7)

fit.glm <- train(admit~ rank + gre + gpa , data=dataset, method="glm", metric=metric, trControl=control)

set.seed(7)

fit.svm <- train(admit~ rank + gre + gpa , data=dataset, method="svmRadial", metric=metric, trControl=control)

# summarize accuracy of models

results <- resamples(list(lrg=fit.glm, svm=fit.svm))

summary(results)

dotplot(results)

# summarize Best Model

print(fit.glm)

predictions <- predict(fit.glm, validation)

confusionMatrix(predictions, validation$admit)

#Descriptive:

# Categorize the average of grade point into High, Medium, and Low (with admission probability percentages) and plot it on a point chart.

#Cross grid for admission variables with GRE Categorization is shown below:

gre\_Table <- dataset %>%

mutate(grep\_prob =

case\_when(gre <440 ~ "Low",

gre <580 ~ "Medium",

gre >=580 ~ "High"))

# Print table

df <- gre\_Table %>%

group\_by(gre\_prob) %>%

summarise(counts = n())

df

# Printpoint chart

ggplot(df, aes(x=grep\_prob, y=counts, color=as.factor(grep\_prob), shape=as.factor(grep\_prob))) +

geom\_linerange(

aes(x = grep\_prob, ymin = 0, ymax = counts),

color = "lightgray", size = 1.5

)+

geom\_point(aes(color = grep\_prob), size = 2)+

ggpubr::color\_palette("jco")+

theme\_pubclean()

# Print bar chart

ggplot(df, aes(x = grep\_prob, y = counts)) +

geom\_bar(fill = "#0073C2FF", stat = "identity") +

geom\_text(aes(label = counts), vjust = -0.3) +

theme\_pubclean()

# Analysis Tasks for admission

# Summary: Most accuarate modle logistic regression with 61% accuracy with feautere

# importances GRE,Rank and GPA

# Descriptive for admission

# GRE scpore plays a vital role in admission . we can see same by data and graphs