Assignment 4: Modelling Mall Customers Data

Sowmya

2022-10-06

# Section 1

The rise in the use of online stores has created a concern for the brick and motor stores and by extension the malls. Mall owners find themselves at the risk of losing tenants if they are unlikely to attract more people to the malls and help the brick and motor stores that rent spaces in their malls. In this project, linear regression modelling is used in the prediction of the shopping score of customers that visits a mall. the model uses Age, Gender and Annual Income of the customers to predict their shopping score. Through the model, the mall owners can target marketing and advertising their malls based on information from the interpretation.

# Section 2

The source of the data used for the linear regression model in this project is <https://www.kaggle.com/datasets/shwetabh123/mall-customers>.

# Section 3

Data loading for the project used the following code

#Loading Data  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --  
## v ggplot2 3.3.6 v purrr 0.3.4   
## v tibble 3.1.8 v dplyr 1.0.10  
## v tidyr 1.2.1 v stringr 1.4.1   
## v readr 2.1.2 v forcats 0.5.2   
## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

Data <- read\_csv("Mall\_Customers.csv")

## Rows: 200 Columns: 5  
## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (2): CustomerID, Genre  
## dbl (3): Age, Annual Income (k$), Spending Score (1-100)  
##   
## 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.

# Section 4

The data summary showed that no NAs were present in any of the data variables. The summary also showed that there were 5 variables hence 5 columns; CustomerID, Genre, Age, Annual Income (k$) and Spending Score (1-100). From the length of the CustomerID variable, the data has 200 rows. The Age variable is shown to have a mean of 38.85 (approx. 39) with the minimum age = 18 and the maximum age = 70. The Annual Income variable (in thousands of dollar) has a mean of 60.56 with the minimum income = 15 and maximum income = 137. Also, the Spending Score variable is shown to have a mean of 50.20 with a minimum score = 1.00 and maximum score = 99.00. Frequency summary for the gender variable shows that data was collected for 112 females and 88 males.

#Summary  
summary(Data)

## CustomerID Genre Age Annual Income (k$)  
## Length:200 Length:200 Min. :18.00 Min. : 15.00   
## Class :character Class :character 1st Qu.:28.75 1st Qu.: 41.50   
## Mode :character Mode :character Median :36.00 Median : 61.50   
## Mean :38.85 Mean : 60.56   
## 3rd Qu.:49.00 3rd Qu.: 78.00   
## Max. :70.00 Max. :137.00   
## Spending Score (1-100)  
## Min. : 1.00   
## 1st Qu.:34.75   
## Median :50.00   
## Mean :50.20   
## 3rd Qu.:73.00   
## Max. :99.00

table(Data$Genre)

##   
## Female Male   
## 112 88

# Section 5

The data was generally clean with possible data cleaning including removal of missing values for particularly the Annual Income variable where the customers might have declined to provide information. The missing entries and data structures represents among the main problems associated with raw data. the data preparation coinducted in this project involved the removal of the CustomerID variable, which will not be used in the modelling. The columns were also renamed in the data to correspond to the actual information and simplify for calling in the model. Finally, the data was separated into training and testing sets in the ratio of 80:20 for the modelling.

#Data Preparation  
#Removing the customer ID variable  
Data <- Data %>% select(-CustomerID)  
#Renaming variables   
colnames(Data) <- c("Gender", "Age", "Annual.Income", "Spending.Score")  
#Separating data into training and testing set  
set.seed(200)  
ind.train <- sample(1:200, 0.8\*nrow(Data), replace = F)   
set.train <- Data[ind.train, ]  
set.test <- Data[-ind.train, ]

# Section 6

The linear regression modelling is a useful tool for the prediction of numerical outcomes in business problems. The model uses factors that are thought or identified to have influence on the numerical outcome for the prediction. In this project the spending score in the dependent variable (numerical outcome) while Gender, Age and Annual Income are the factors of interest and independent variable in the model.

# Section 7

The modelling process and the model output are provided as follow below. The results show that predictive model takes the following form of an equation

#Linear Regression  
fit.linear <- lm(Spending.Score~., data = set.train)  
summary(fit.linear)

##   
## Call:  
## lm(formula = Spending.Score ~ ., data = set.train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -55.418 -17.401 1.225 17.954 45.584   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 72.39870 7.29154 9.929 < 2e-16 \*\*\*  
## GenderMale -1.86839 3.86501 -0.483 0.629482   
## Age -0.52835 0.13524 -3.907 0.000139 \*\*\*  
## Annual.Income -0.02583 0.07275 -0.355 0.722973   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 24.2 on 156 degrees of freedom  
## Multiple R-squared: 0.09237, Adjusted R-squared: 0.07492   
## F-statistic: 5.292 on 3 and 156 DF, p-value: 0.001679

The check on the model performance above shows that the Adjusted R-Squared = 0.07492 implying that the model fitness is poor at 7.492%. The performance of the model prediction on the testing set presented below was low with the correlation between the observations and prediction = 0.1765335

#Predictions  
Predict.linear <- predict(fit.linear, newdata = set.test[,-4])  
#Predictive Performance  
library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

R2(pred = Predict.linear, obs = set.test$Spending.Score)

## [1] 0.1765335

# Section 8

The model visualization are not available for the linear regression model. A possible visualization would be a scatterplot of the Annual Income against the Spending Score with Age used as weight for the color of the data points. Two regression lines could be included in the plot for each gender. This would provide a means for representing the model on a visualization.