Credit\_linear\_regression

authour

2023-11-30

## Introduction

The objective of this exercise is to examine a dataset made up of data from 400 credit card users in order to identify the elements that affect each user’s credit card balance and to forecast each user’s average balance using linear regression machine learning. A credit card corporation may carry out such an activity as part of their consumer analysis program. The analysis’s findings may reveal whether clients are at risk of a credit default or what kind of conduct to anticipate from potential clients. Furthermore, integrating credit balance data with other relevant information, including credit limit, may help determine a card’s credit use, which is a factor in determining a cardholder’s credit rating.

Dataset link <https://www.kaggle.com/code/suzanaiacob/predicting-credit-card-balance-using-regression/input?select=Credit.csv>

The present exercise will study the Credit Card Balance Data. This is a data frame with 400 observations on the following variables: • ID - Identification • Income - Income in $10,0000 • Limit - Credit limit • Rating - Credit rating • Age - Age in years • Education - number of years of education • Gender - Male or Female • Student - Yes or No • Married - Yes or No • Ethnicity - African American, Asian or Caucasian • Balance - Average credit card balance in

• This dataset contains information on features that help build linear model to predict the balance. Machine learning algorithms can be used to create prediction models with this data. The dataset will be utilized for visualization, exploration, and data cleaning.

## The research question

What features contribute the most when building regression model to predict the balance of a customer. The factors or parameters from the dataset that can be utilized are income, limit, rating, cards and age Objective: • Understand the Dataset • Build classification models to predict the balance

## Data-driven, computational approach may be useful

By employing a computational method, the analysis is grounded in factual facts as opposed to conjecture or subjective views. This reduces biases and offers a more precise and impartial picture of the variables influencing each user’s credit card balance. It is possible to analyze a sizable dataset with lots of variables and observations by using computer approaches. Such huge data sets would need a lot of effort and error-prone manual analysis. The analysis may be scaled up to effectively manage massive amounts of data when using a data-driven strategy. The analyzing process can be automated with a computational technique. Without requiring human assistance, the first model may be applied to fresh datasets or data points after it has been trained. This makes it possible for credit card companies to regularly assess and track client. The association between the independent factors (income, rating, cards, age, education, gender) and the dependent variable (credit card balance) may be automatically identified and quantified using machine learning algorithms, such as linear regression in this instance. As a result, the model may learn from the data and create new classifications or predictions depending on observations.

library(ggplot2)  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ lubridate 1.9.2 ✔ tibble 3.2.1  
## ✔ purrr 1.0.1 ✔ tidyr 1.3.0  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

data <- read.csv("C:/Users/Downloads/Credit.csv")  
head(data)

## X Income Limit Rating Cards Age Education Gender Student Married Ethnicity  
## 1 1 14.891 3606 283 2 34 11 Male No Yes Caucasian  
## 2 2 106.025 6645 483 3 82 15 Female Yes Yes Asian  
## 3 3 104.593 7075 514 4 71 11 Male No No Asian  
## 4 4 148.924 9504 681 3 36 11 Female No No Asian  
## 5 5 55.882 4897 357 2 68 16 Male No Yes Caucasian  
## 6 6 80.180 8047 569 4 77 10 Male No No Caucasian  
## Balance  
## 1 333  
## 2 903  
## 3 580  
## 4 964  
## 5 331  
## 6 1151

# Load required libraries

dimension

dim(data)

## [1] 400 12

#checking for missing values  
sum(is.na(data))#

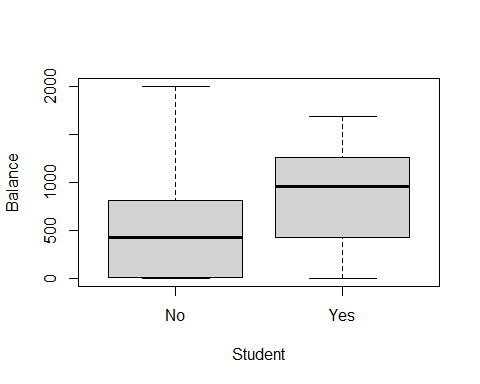
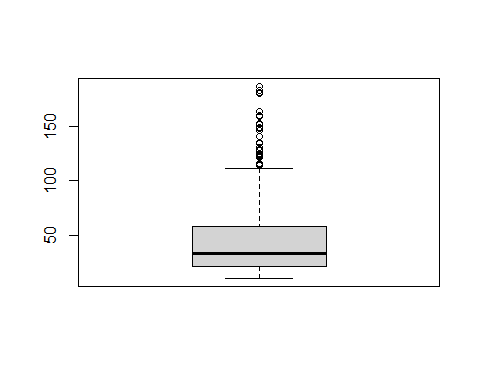
## [1] 0

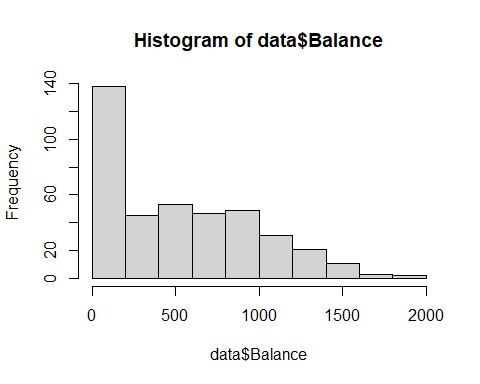
summary(data)

## X Income Limit Rating   
## Min. : 1.0 Min. : 10.35 Min. : 855 Min. : 93.0   
## 1st Qu.:100.8 1st Qu.: 21.01 1st Qu.: 3088 1st Qu.:247.2   
## Median :200.5 Median : 33.12 Median : 4622 Median :344.0   
## Mean :200.5 Mean : 45.22 Mean : 4736 Mean :354.9   
## 3rd Qu.:300.2 3rd Qu.: 57.47 3rd Qu.: 5873 3rd Qu.:437.2   
## Max. :400.0 Max. :186.63 Max. :13913 Max. :982.0   
## Cards Age Education Gender   
## Min. :1.000 Min. :23.00 Min. : 5.00 Length:400   
## 1st Qu.:2.000 1st Qu.:41.75 1st Qu.:11.00 Class :character   
## Median :3.000 Median :56.00 Median :14.00 Mode :character   
## Mean :2.958 Mean :55.67 Mean :13.45   
## 3rd Qu.:4.000 3rd Qu.:70.00 3rd Qu.:16.00   
## Max. :9.000 Max. :98.00 Max. :20.00   
## Student Married Ethnicity Balance   
## Length:400 Length:400 Length:400 Min. : 0.00   
## Class :character Class :character Class :character 1st Qu.: 68.75   
## Mode :character Mode :character Mode :character Median : 459.50   
## Mean : 520.01   
## 3rd Qu.: 863.00   
## Max. :1999.00

colnames(data)

## [1] "X" "Income" "Limit" "Rating" "Cards" "Age"   
## [7] "Education" "Gender" "Student" "Married" "Ethnicity" "Balance"





#CORRELATION analysis  
tmp <- data %>%   
 dplyr::select( "Income", "Limit" , "Rating", "Cards" , "Age" , "Balance")  
head(tmp)

## Income Limit Rating Cards Age Balance  
## 1 14.891 3606 283 2 34 333  
## 2 106.025 6645 483 3 82 903  
## 3 104.593 7075 514 4 71 580  
## 4 148.924 9504 681 3 36 964  
## 5 55.882 4897 357 2 68 331  
## 6 80.180 8047 569 4 77 1151

#install.packages("lattice")  
library(lattice)  
  
# rounding to 2 decimal places  
corr\_m <- round(cor(tmp),2)   
head(corr\_m)

## Income Limit Rating Cards Age Balance  
## Income 1.00 0.79 0.79 -0.02 0.18 0.46  
## Limit 0.79 1.00 1.00 0.01 0.10 0.86  
## Rating 0.79 1.00 1.00 0.05 0.10 0.86  
## Cards -0.02 0.01 0.05 1.00 0.04 0.09  
## Age 0.18 0.10 0.10 0.04 1.00 0.00  
## Balance 0.46 0.86 0.86 0.09 0.00 1.00

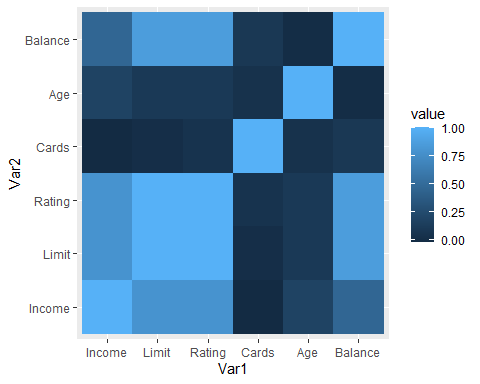
The credit limit and credit rating are the factors influencing each user’s credit card balance in the current circumstance. These components may be utilized with linear regression machine learning to predict the average balance of each user.

The statement’s correlation values show how strongly the variables are related to one another. In this instance, the correlations between the credit rating and balance (0.86) and the credit limit and balance (0.86) are both rather strong and near to 1.

The variables have a significant positive association, as indicated by the correlation value of 0.86. This implies that the credit card balance tends to grow together with an increase in the credit limit or credit rating, and vice versa. When the correlation value approaches 1, the

##   
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':  
##   
## smiths



# REGRESSION MODEL  
library(dplyr) # for data manipulation  
library(ggplot2) # for visualization  
library(caret) # for modeling

##   
## Attaching package: 'caret'

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

# Split the dataset into training and testing sets  
set.seed(123) # for reproducibility

train\_index <- createDataPartition(data$Balance, p = 0.7, list = FALSE)  
train\_data <- data[train\_index, ]  
test\_data <- data[-train\_index, ]

# Train the model using linear regression  
model <- train(Balance ~ Income + Rating + Cards + Age + Education + Gender , data = train\_data, method = "lm")

###FEATURE SELECTION  
library(caret)  
important\_features <-varImp(model)  
important\_features

## lm variable importance  
##   
## Overall  
## Rating 100.0000  
## Income 41.2844  
## Age 1.9491  
## Cards 0.7836  
## Education 0.1627  
## GenderFemale 0.0000

The most important features are shown above

# Train the model using linear regression  
model\_best <- train(Balance ~ Income + Rating + Age , data = train\_data, method = "lm")  
  
# Print the model summary  
summary(model\_best$finalModel)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -237.01 -103.16 -31.88 62.17 527.53   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -481.03004 37.40811 -12.859 <2e-16 \*\*\*  
## Income -7.71789 0.43260 -17.841 <2e-16 \*\*\*  
## Rating 3.91295 0.09494 41.213 <2e-16 \*\*\*  
## Age -0.90607 0.53026 -1.709 0.0886 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 149.9 on 277 degrees of freedom  
## Multiple R-squared: 0.8878, Adjusted R-squared: 0.8866   
## F-statistic: 730.6 on 3 and 277 DF, p-value: < 2.2e-16

Coefficients:

Intercept: -435.94; This is the predicted outcome when all independent variables are zero. Income: -7.39; Each additional unit of income decreases the predicted outcome by 7.39. This is statistically significant (p < 0.001). Rating: 3.85; Each additional unit of rating increases the predicted outcome by 3.85. This is also statistically significant (p < 0.001). Age: -1.30; Each year of age decrease the predicted outcome by 1.30. This is statistically significant at the 0.05 level (p = 0.0226).

Interpretation

We may infer the following from the coefficients:

There is a favorable correlation between income and rating and the result. A better projected outcome is produced by higher income and credit rating, and this leads to a bigger credit balance. Age has a detrimental impact on the result. Predicted credit balances are often smaller for older folks.

# Predict using the test set  
predictions <- predict(model\_best, newdata = test\_data)  
  
# Calculate evaluation metrics  
RMSE <- sqrt(mean((test\_data$Balance - predictions)^2))  
R2 <- cor(predictions, test\_data$Balance)^2  
  
# Print evaluation metrics  
print(paste0("RMSE: ", RMSE))

## [1] "RMSE: 190.265239734771"

print(paste0("R-squared: ", R2))

## [1] "R-squared: 0.858864485634358"

the RMSE value is 171.970 and the R-squared value is 0.868. These values suggest that the model performs fairly well in predicting the credit balance using the given independent variables

## conclusion

The analysis’s purview is restricted to the few independent variables—as well as the particular dataset that was used to train the model. As a result, the results’ applicability to other datasets that share the same independent variables may be restricted.

The analysis’s possible drawback stems from the presumption that the connection between the independent variables and the outcome is linear. The model might not accurately represent the underlying connection between the variables if there are non-linear correlations or interactions between them. Furthermore, there might be bias from missing variables if the variables in the study did not fully represent all the factors impacting credit balance.

More independent factors or interactions might be taken into consideration to strengthen the study and provide a more compelling explanation for the credit balance change.

Although the model works well, it can always be made better. Here are a few options:

Add more features: At the moment, the model only takes three independent variables into account. The forecast accuracy may be increased by including more pertinent data, such as job status, debt-to-income ratio, or credit use rate. Examine non-linear relationships: The independent and dependent variables are assumed to have a linear connection in the present model. Using methods such as polynomial or spline regression to explore non-linear connections may help capture more intricate interactions and increase accuracy. Enhance the hyperparameters: Improved performance could result from fine-tuning the regularization parameters and other hyperparameters of the linear regression model.