Sang Hwa Lee

INST 354

December 11, 2022

Exercise 4

1. How many rows and columns does the bankchurn dataset have? What are the variables in the bankchurn dataset? Which variables, if any, have NA values in them?

   Rows = 9900,   Columns = 14,   There are no NA values in the data.

2. Produce some descriptive statistics for the 'CreditScore' variable. Specifically, what is the minimum, maximum, and mean credit score.

   Min = 350,   Max = 850,   Mean = 650.3768

3. Identify two variables in the dataset that you think could plausibly be associated with credit score.   Explain your thinking. (2-3 sentences) Note: this question does not require any data analysis.

   Except for data with binary flags that can make errors in data analysis, the variables that can affect credit scores are estimated salary and tenure.

   Tenure that provides accurate information and performance related to long-term credit. Also, estimated salary that can determine the likelihood of insolvency through constant earnings can affect credit scores.

4. Suppose you are asked to determine whether customer age is associated with higher credit scores. Produce some descriptive statistics for the 'Age' variable. Specifically, what is the minimum, maximum, and mean customer age.

   Min =18,   Max = 92,   Mean = 38.92657

5. Is there a significant (p<0.05) relationship between age and credit score? Compute a correlation and its associated p-value to determine this.

   There seems to be a negative relationship between the two variables (r = -0.002391807)

   and the p-value = 0.8119. The p-value is greater than 0.05, so the correlation is not statistically significant.

6. Is there a significant (p<0.05) relationship between age and credit score, after controlling for customer "Geography" and "Gender"? Run a regression model predicting credit score to determine this. In your answer, focus on whether the coefficient for age is positive or negative and whether its p-value is less than 0.05.

   The coefficient for age is negative.   Age = - 0.02595 and P-value > 0.05

   Therefore, it looks like there is not a significant relationship between the age and the credit score when the "geography" and "gender" are controlled.

7. Would you feel confident using this predictive model to make business decisions? Why or why not? Your answer's foci should include ethical issues and challenges. (a brief paragraph)

   There is a chance that this predictive model may have a bias. First, there is the problem of binary classification. Binary classification may count multiple actions as once. This data set has several binary flags, such as HasCrCard, IsActiveMember, Exited. Also, when allocating finite resources, the privileged class is favored compared to the more vulnerable. Furthermore, the variables of geography and gender may cause bias. Therefore, I would not feel confident in using this predictive model to make business decisions.

R studio code

```r
#Before starting the lab
bankchurn <- read.csv('C:/Users/sangh/Downloads/bankchurn_ex4.csv', header = TRUE)

# 1
nrow(bankchurn)
ncol(bankchurn)
is.na(bankchurn)

# 2
min(bankchurn$CreditScores)
max(bankchurn$CreditScores)
mean(bankchurn$CreditScores)

# 4
min(bankchurn$Age)
max(bankchurn$Age)
mean(bankchurn$Age)

# 5
cor(bankchurn$Age, bankchurn$CreditScores)
cor.test(bankchurn$Age, bankchurn$CreditScores)

# 6
mod1 <- lm(bankchurn$CreditScores ~ bankchurn$Age + bankchurn$Geographies + bankchurn$Gender)
summary(mod1)
```