Graded Project – Inferential Statistics

70 Marks

PART -I (Probability)

Problem Statement:

Investment Advisors agree that near retirees, defined as people aged 55 to 65, should have balanced portfolios. Most advisors suggest that the near –retirees have no more than 50% of their investments in stocks. However, during the huge decline in the stock market in 2008, 22% of near retirees had 90% or more of their investments in stocks.

Suppose you have a random sample of 10 people who have been labeled as near retirees in 2008. Let X (near retirees) be a random variable which follows Binomial Distribution with the parameters n = 10 and p = 0.22.

Domain: Finance, Investment

Steps and tasks (20 points)

1. What is the probability that during 2008 zero near retirees had 90% or more of their investments in stocks? (5 point)
2. What is the probability that during 2008 exactly one near retiree had 90% or more of his investments in stocks? (5 point)
3. What is the probability that during 2008 two or fewer near retirees had 90% or more of their investment in stocks? (5 point)
4. What is the probability that during 2008 three or more near retirees had 90% or more of their investments in stocks? (5 points)

PART -II - Hypothesis Testing (50 Marks)

Problem Statement: The card payments data is published by the Reserve Bank of India on a monthly basis. The statistics cover the methods of payment used in retail transactions and ATM transactions in India. It constitutes payments via debit cards, credit cards, ATMs etc. It can be used to analyze the trend and adaptation of digital payment over the years.

Using basic Python functions, we want to analyse the pattern of usage of credit card and debit card over the years. We will also see the trend of digital payment by various banks over the years. We will analyze a few patterns like how many ATMs and PoS Machines have been installed over the years. Which bank has topped in terms of deploying ATMs. What is the

percentage of credit and debit cards, etc.

Domain: Banking

Data Description: payment_data.csv - The data set contains information of monthly payments made by various bank ATMs and PoS terminals. This data is published by the Reserve Bank of India. This data has information from Apr'2011 to Aug'2019

Attribute Information:

month: month of the year

year: year

month_number : month number in integers

start_date : start date

end_date : end date for the month

bank_name : Bank name

no_atms_on_site : Number of ATMs deployed on site by the bank

no_atms_off_site : Number of ATMs deployed off site by the bank.

no_pos_on_line : Number of POS deployed online by the bank

no_pos_off_line : Number of POS deployed offline by the bank

no_credit_cards: Total number of credit cards issued outstanding (after adjusting the number of cards withdrawn/cancelled).

no_credit_card_atm_txn : Total number of financial transactions done by the credit card issued by the bank at ATMs

no_credit_card_pos_txn : Total number of financial transactions done by the credit card issued by the bank at POS terminals

no_credit_card_atm_txn_value_in_mn : Total value of financial transactions done by the credit card issued by the bank at ATMs

no_credit_card_pos_txn_value_in_mn : Total value of financial transactions done by the credit card issued by the bank at POS terminals

no_debit_cards : Total number of debit cards issued outstanding (after adjusting the number of cards withdrawn/cancelled).

no_debit_card_atm_txn : Total number of financial transactions done by the debit card issued by the bank at ATMs

no_debit_card_pos_txn : Total number of financial transactions done by the debit card issued by the bank at POS terminals

no_debit_card_atm_txn_value_in_mn : Total value of financial transactions done by the

debit card issued by the bank at ATMs

no_debit_card_pos_txn_value_in_mn : Total value of financial transactions done by the debit card issued by the bank at POS terminals

Steps and Tasks

Data Understanding (6 points)

1. Read the data set and check shape and info and get familiar with the data. (2 point)
2. Check the summary statistics of the dataframe and comment on your findings. (2 point)
3. Check for null values and impute them with appropriate values. (2 point)

Descriptive statistics of individual features – Use Plots if necessary (15 points)

4. Print the number of unique banks in the data. Display the top 10 and bottom 10 banks. (5 points)
    a. consider the same bank with different spellings as different banks to avoid Data Cleansing.
    b. Please clean the data and display the count again
5. Check the trend of features w.r.t year. You can use groupby to group the yearly data and display the yearly mean value for each feature (any two features). (5 points)
6. Which feature has shown the highest growth over the years.
    a. Check for all features exclude transition related fetures (number of atm, offsite, pos line, etc)
    b. Check for the year on year growth for transactions related features as they are in millions. (5 Points)

Descriptive statistics of multiple features - Use plots if (15 points)

7. List the top 5 banks having the highest number of ATMs deployed on site (no_atms_on_site). (3 points)
    a. Count the number of ATMs against each back
    b. Display the percentage of the “number of ATMs deployed on site” by banks using appropriate plots.
8. Create a new column having total number of transactions
    a. Create a new column both credit card and debit card transactions from atm and pos
    b. list top 5 banks having the highest number of transactions. (3 points)
9. Plot the trend of ‘total transactions” across years using appropriate plots. (3 point)
10. Drop these features for the next question. cols_to _drop = ['month', 'year', 'month_number', 'start_date', 'end_date', 'bank_name'](3 point)
11. Check correlation matrix and display heatmap. Comment on your findings. (3 points)

Hypothesis Testing (9 points)

Note:- Consider a significance level of 5%

12. Is there any significant difference between the mean values of the number of debit

cards (no_debit_cards) and the number of credit cards (no_credit_cards) over the years? (3 points) (hint:- Create a sample having the mean of two features grouped by year and then apply a relevant test).

a. Write the null hypothesis and alternative hypothesis for the above case.
b. Define the test that should be used and the significance level.
c. Perform the hypothesis testing and report your conclusion.

13. Is there any significant difference among the means of the following transactions over the years? (3 points)
    i. no_credit_card_atm_txn_value_in_mn
    ii. no_credit_card_pos_txn_value_in_mn
    iii. no_debit_card_atm_txn_value_in_mn
    a. Write the null hypothesis and alternative hypothesis for the above case.
    b. Define the test that should be used and the significance level.
    c. Perform the hypothesis testing and report your conclusion.
14. Is there any significant difference between the mean values of **no_atms_on_site** of two banks (BANK OF BARODA and HDFC BANK LTD) over the years? (3 points)
    a. Write the null hypothesis and alternative hypothesis for the above case.
    b. Define the test that should be used and the significance level.
    c. Perform the hypothesis testing and report your conclusion.

Business Report: Comments, Inferences and Suggestions (5 Marks)

15. Business report, Inferences and observations for all the Tasks
    a. Each step needs to be executed clearly separately gradually. Also ensure you provided proper comments at each step and provide respective inferences or observations
    b. Please note the hypothesis inferences, Observations and Conclusion should be available in both the workbook (.ipnyb) and the report clearly.
    c. All the answers to the questions to be submitted in a sequential manner as part of the business repot
    d. The Business report should include a detailed explanation of the approach used, insights, inferences, all outputs like graphs, tables, etc. The Level of detail and explanation in business report should be deeper
    e. Your report should not be filled with formulas. Only use important metrics or values or charts obtained from each step of analysis