Assignment 1 STA4102

Due: Oct 03, 2022

Put all working out and code into a single pdf file. Partial credit is awarded where an incorrect answer is alongside an explanation showing the working out.

Please use either R, SAS, Python, or Julia for the question(s) requiring the use of software.

Walmart Sales Data in file 'WalmartSalesData.csv'

## Walmart data (5 points)

There are 12 columns in this dataset.

1. Store: Store number
2. Date: Week
3. Temperature: Average temperature in the region
4. Fuel_Price: Cost of fuel in the region
5. MarkDown1: Anonymized data related to promotional markdowns that Walmart is running.
6. MarkDown2: Anonymized data related to promotional markdowns that Walmart is running.
7. MarkDown3: Anonymized data related to promotional markdowns that Walmart is running.
8. MarkDown4: Anonymized data related to promotional markdowns that Walmart is running.
9. MarkDown5: Anonymized data related to promotional markdowns that Walmart is running.
10. CPI: The consumer price index
11. Unemployment: The unemployment rate
12. IsHoliday: Whether the week is a special holiday week

Questions:

1. Load the data
2. Print the first and last six rows of the dataset
3. Find the dimension of the dataset

4. Find the total number of missing values
5. Find the missing values for each column
6. Visualize this data in any way you choose by producing 3 separate figures
7. Remove the missing values or perform **imputations** where you can replace missing values with your chosen approach (+2 points for imputations)
8. Show that there are no more missing values

## Credit Card Data (13 points)

Credit Card Data in file 'CreditCardData.csv'
There are 20 columns in this dataset.

1. Client ID
2. Gender: Gender (1=Male, 0=Female)
3. Own_car: Does the client own a car? (1=Yes, 0=No)
4. Own_property: Does the client own property? (1=Yes, 0=No)
5. Work_phone: Does the client own a work phone? (1=Yes, 0=No)
6. Phone: Does the client own a phone? (1=Yes, 0=No)
7. Email: Does the client have an email address? (1=Yes, 0=No)
8. Unemployed: Is the client unemployed? (1=Yes, 0=No)
9. Num_children: Number of children
10. Num_family: Number of family members
11. Account_length: Number of months credit card has been owned
12. Total_income: Total income (Chinese Yuan)
13. Age: Age in years
14. Years_employed: Number of years employed
15. Income_type
16. Education_type
17. Family_status
18. Housing_type
19. Occupation_type

20. Target: Target (1=high risk, 0=low risk)

(Before delving into further questions, get familiar with the data and check if there are any missing values)

1. Partition the data into train-test (eg 70-30)
2. Using the training dataset produce a decision tree model with the 'Unemployed' value as the dependent and the set of independents as 'Own_property', 'Phone', 'Email', 'Num_children'. Visualize the tree that is fitted. Use the testing dataset and report the accuracy in predicting the Unemployed values.
3. Using the training dataset produce a decision tree model with the 'Age' as the dependent variable and the independents as the 'Total_income', 'Num_children', 'Unemployed', 'Own_property'. Visualize the tree that is fitted. Use the testing dataset and report the RMSE.
4. Produce a decision tree to predict the Target column values of high risk or low risk. Using the accuracy on the test dataset which 4 independent variables (features) do you choose reporting the accuracy for the alternatives list at least 3)?

**Credit contour visualization (2 points)**

Produce a contour plot where the horizontal axis is the Account_length, the vertical axis is the Age and the z-index is the Total_income. The z-index can be produced using a linear model (regression) or with decision trees.