Business Analytics with R – Assignment 1

Name – Amruth Pai Thukaram

Net ID – AXT230147

Date – 17th September 2024

1. Assuming that data mining techniques are to be used in the following cases, identify

whether the task required is supervised or unsupervised learning. (5 points)

a. Deciding whether to issue a loan to an applicant based on demographic and

financial data (with reference to a database of similar data on prior customers).

Supervised

b. In an online bookstore, making recommendations to customers concerning

additional items to buy based on the buying patterns in prior transactions.

Unsupervised

c. Identifying a network data packet as dangerous (virus, hacker attack) based on

comparison to other packets whose threat status is known.

Supervised

d. Identifying segments of similar customers.

Unsupervised

e. Predicting whether a company will go bankrupt based on comparing its financial

data to those of similar bankrupt and nonbankrupt firms.

Supervised

2. Identify whether the following are regression or classification tasks. (10 points)

a. Predicting the price of automobiles based on the features like make, engine-type,

number of doors, fuel-type, etc.

regression

b. Predicting the income of people based on the features like occupation, age,

gender, education level, marital status, etc.

regression

c. Predicting whether income is above or below 50K based on the features like

occupation, age, gender, education level, marital status, etc.

classification

d. Predicting the average life expectancy of different countries based on their GDP,

population, schooling, and health-related metrics.

Regression

e. Predicting whether a customer would cancel their hotel booking or not based on

the features like when the reservation was made, how many rooms were reserved,

how the rooms were reserved, etc.

classification

3. Shipments of Household Appliances: Line Graphs. The file

ApplianceShipments.csv contains the series of quarterly shipments (in millions of

dollars) of US household appliances between 1985 and 1989. (20 points)

a. Create a well-formatted time plot of the data using R.

A line of lines on a white background

Description automatically generated

b. Does there appear to be a quarterly pattern? For a closer view of the patterns,

zoom into the range of 3500–5000 on the y-axis.

A graph on a white background

Description automatically generated

c. Using R, create one chart with four separate lines, one line for each of Q1, Q2, Q3,

and Q4. In R, this can be achieved by generating a data.frame for each quarter Q1,

Q2, Q3, Q4, and then plotting them as separate series on the line graph. Zoom

into the range of 3500–5000 on the y-axis. Does there appear to be a difference

between quarters?

A graph of different colored lines

Description automatically generated

d. Using R, create a line graph of the series at a yearly aggregated level (i.e., the total

shipments in each year).

A graph with a line

Description automatically generated

4. Sales of Riding Mowers: Scatter Plots. A company that manufactures riding

mowers wants to identify the best sales prospects for an intensive sales campaign.

In particular, the manufacturer is interested in classifying households as prospective

owners or nonowners on the basis of Income (in $1000s) and Lot Size (in 1000 ft2).

The marketing expert looked at a random sample of 24 households, given in the file

RidingMowers.csv. (5 points)

a. Using R, create a scatter plot of Lot Size vs. Income, color-coded by the outcome

variable owner/nonowner. Make sure to obtain a well-formatted plot (create legible

labels and a legend, etc.).

A graph of a graph

Description automatically generated with medium confidence

5. Laptop Sales at a London Computer Chain: Bar Charts and Boxplots. The

file LaptopSalesJanuary2008.csv contains data for all sales of laptops at a computer

chain in London in January 2008. This is a subset of the full dataset that includes

data for the entire year. (10 points)

a. Create a bar chart, showing the average retail price by store. Which store has the

highest average? Which has the lowest?

A graph of blue rectangular bars

Description automatically generated with medium confidence

Highest: the highest average retail price is 494.6341 at store post code N17 6QA

Lowest: the lowest average retail price is 481.0063 at store post code W4 3PH

A white background with blue text

Description automatically generated

b. To better compare retail prices across stores, create side-by-side boxplots of

retail price by store. Now compare the prices in the two stores from (a). Does there

seem to be a difference between their price distributions?

A graph with lines and dots

Description automatically generated with medium confidence

This is a better comparison, looks like the we can differentiate in this graph better but they have a very similar average price