Southern New Hampshire University

DAT 610 Module Four Exercise Guidelines and Rubric

Overview: A robust operational risk program includes an emphasis on the proactive identification of losses as well as a data capture for those that have already been incurred.

Prompt: Company XYZ has tasked you with an evaluation of analytic tools to bolster its operational loss program—especially the identification of losses to fraudulent bodily injury claims. Your initial task is to improve Company XYZ's fraud identification efforts by validating the advanced analytic functionality in R Studio, an open-source statistical and visualization software package.

Use R Studio to apply the PRIDIT [1] scoring model to the Auto Accident Personal Injury Claims[2] document, located in the Assignment Guidelines and Rubrics section of the course.

1. Describe the PRIDIT scoring methodology. In your answer, reference a process to compare the PRIDIT scoring method against other scoring methods. The Module Four Exercise Introduction and Additional Resources document should help you with this.

2. Inspect the RIDIT-ized variables in the claims data file. Confirm that each RIDIT transformation results in appropriate directionality for the PRIDIT scoring method.

   This question is composed of two parts. To start thinking about this problem, let's think about what we do know and what we don't.

   ✔ We have the Auto Accident Personal Injury Claims file with 20 variables that correlate to a suspicion score and 20 RIDIT transformations of these variables.

   ✘ We don't know how the original scores were tabulated, so we have no idea about direction. The best way to verify directionality would be a series of scatterplots of the raw variable data versus its PRIDIT score using a scatterplot matrix. This will tell you visually.

   Part One: Inspect the RIDIT-ized variables in the claims data file using R Studio.

   Hint: It might be easier to save your .csv file to "DAT610" or something simple for reading it into R Studio and referencing it in your code.

   Step 1: Read in your data file. Remember to set your working directory to your Course Content folder or wherever you saved your data file. Suggested code:

```
setwd("U:/")
auto=read.csv("DAT 610 Auto Accident Personal Injury Claims.csv",header=TRUE, sep=",")
```

Step 2: Set your variable myData.
Suggested code:

```
myData=auto[1:502,6:25]
```

Step 3: Set variable rid to inspect the RIDIT-ized variables.
Suggested code:

```
rid=data.frame("RIDIT" = cbind("_01"= 2*(0 + 0.5*table(myData$IND_01)[1]/502) - 1,"i2"=2*(table(myData$IND_01)[1]/502 +
0.5*table(myData$IND_01)[2]/502) - 1,"i5"=2*(table(myData$IND_02)[1]/502 +
table(myData$IND_02)[2]/502+table(myData$IND_02)[3]/502+table(myData$IND_02)[4]/502+0.5*table(myData$IND_02)[5]/502) - 1))
```

Step 3 formulas explained:

- $P(X = Xi)$ is the probability of number 1 or 2 or ... 5 occurring; table counts the occurrences of each number in the column.
- Formula => $2 * (P(X<Xi) + 0.5*P(X=Xi)) – 1$
- The code above gives the calculations for the RIDIT where i =1, 2, or 5.

Step 4: Type in rid, hit enter, and see your results.

Part Two: Confirm that each RIDIT transformation results in appropriate directionality for the PRIDIT scoring method.
Determining directionality is key to solving this problem. Directionality tells you the effect of a risk variable and whether that variable makes your risk increase or decrease.

Step 1: Generate scatterplot matrix of raw variable data versus its RIDIT score.

3. Perform a principal components analysis of the RIDIT-ized transformations of the risk identification variables in the claims file. Use the R princomp function and report the summary and scores.

Step 1: Set variable myRidit.
Suggested code:

```
myRidit=auto[1:502,26:45]
```

Step 2: Perform a principal components analysis on myRidit using the princomp function:
Suggested code:

```
MyRiditPCA=princomp(~ ., data = myRidit, cor = TRUE)
```

Step 3: View summary of analysis.

Suggested code:

```
summary(myRiditPCA)
```

4. Prepare three graphical visualization of the results of the PRIDIT scoring method to the claims file.

   A. Produce an R screeplot of the PCA analysis and report on the identification of which IND components are needed to summarize the data.

   B. Produce a variables factor map using the FactoMineR packages and report which IND variables show strong correlation.

   - Here, it is important that you are using R Studio.
   - Suggested code:

```
install.packages("FactoMineR")
source("http://factominer.free.fr/install-facto.r")
library(FactoMineR)
#initiate FactoMineR window with commands above
#select dataset: auto
#under rightmost FactoMineR menu, select the second item: PCA
#select all IND vars
#leave all options at default and click OK
#Variables factor map graph is produced
#OR... do this:
res.pca = PCA(auto[,6:25], scale.unit=TRUE, ncp=5, graph=T)
#just select 1 to 20 variables...
#same results as using the GUI and selecting the specific vars.
```

   C. Finally, produce a scatterplot of PCA first component scores versus the claim suspicion scores. Report the significance of any relationships shown in the chart and their potential use for fraud detection.

   - See page 10 of this document.
   - Suggested code:

```
#produce the scatterplots with ellipses
#change claim suspicion score into a factor variable
#we have this problem often, so databank this as.factor technique:
auto$CS2 <- as.factor(auto$CLAIM_SUSPICION_SCORE)
myData2 <- auto[c(6:25, 46)]
res.pca3 = PCA(myData2[1:21], scale.unit=TRUE, ncp=5, quali.sup=21, graph=T)
plotellipses(res.pca3,21)
```

Southern New Hampshire University

What does this visualization tell us? What dimension is associated with higher suspicion? What does this mean?

_____

[1.] Francis, L. (1997, January 1). *Advanced unsupervised learning methods applied to property-casualty databases*. Retrieved from cas.confex.com

[2.] Auto Accident Personal Injury Claims file definitions:

- **Claim Number:** The data were generated to resemble a claims file. Each record represents a bodily injury claim with its own claim number.
- **Policy ID:** Policy ID reflects a number assigned to an auto policy to uniquely identify it.
- **Claim Amount:** The amount claimed against policy benefits.
- **Paid Amount:** The actual amount paid on a given claim.
- **Claim Suspicion Score:** An assessment of the degree of suspicion that is attached to a claim made by a professional adjuster, where 1 through 5 is an assessment range from not suspicious to very suspicious.
- **IND_01-IND_20:** These are indicative of 20 metrics/variables that correlate to the suspicion score. They could be items such as severity of sprain to neck, damage to auto, damage to other property, etc. The indicative metrics (IND) are given in the form of scores ranging from 1 to 5.
- **RIDIT_01-RIDIT_20:** These are the RIDIT-transformed versions of each of the raw metrics.

Guidelines for Submission: This submission must be 1 to 2 pages in length and must use double spacing, 12-point Times New Roman font, and one-inch margins. Citations must use APA format.

Rubric

Southern New Hampshire University

Critical Elements	Exemplary (100%)	Proficient (90%)	Needs Improvement (70%)	Not Evident (0%)	Value
Description of PRIDIT Scoring Method	Meets "Proficient" criteria and references a process to compare the PRIDIT scoring method against other scoring methods	Describes the PRIDIT scoring method specifically in terms that relate to loss data sources other than just the claims file	Describes the PRIDIT scoring method specifically in terms of the claims file	Does not describe the PRIDIT scoring method	25
Evaluation of RIDIT-ized Transformations	Meets "Proficient" criteria and references a process to enhance the claims file with other RIDIT-ized variables	Describes the RIDIT-ized transformation of the claims file variables; describes the cumulative distribution used in RIDIT computation	Describes the RIDIT-ized transformations of the claims file variables	Does not describe the RIDIT-ized claims file variable transformations	25
Core Claims File Observations	Meets "Proficient" criteria and references a process to apply PC scoring to future claims	PCA scores the claims file, reports the first PC derived from the RIDIT-ized claims file, and gives results of the scoring	PCA scores the claims file and reports the first PC derived from the RIDIT-ized claims file but does not give results of the scoring	Does not report PCA scoring summary of the claims file	25
Visualization of Results	Meets "Proficient" criteria and references a process to incorporate other features into the visualization	Creates a visualization that displays highest scoring claims in the claims file	Creates a visualization in which highest scoring claims are not evident	Does not create a visualization	25
				Total	**100%**