**FUNDAÇÃO GETÚLIO VARGAS**

**Business Analytics – Big Data**

**Prof João Fernando Serrajordia Rocha de Mello**

**Disciplina: Modelagem Estatística Avançada – Trabalho 1 (individual)**

**Aluno: Alexandre Vasconcelos Lima**

**Questão 1:** Marque a alternativa que não corresponde a uma liberdade referente à definição de software livre:

A. ( ) Liberdade para executar o programa para qualquer propósito.

B. ( ) Liberdade de estudar e alterar o programa para que ele faça o que você deseja.

|  |  |
| --- | --- |
|  |  |

C. ( ) Liberdade de redistribuir e fazer cópias de modo que você possa ajudar ao próximo.

|  |  |
| --- | --- |
|  |  |

D. ( ) Liberdade de melhorar o programa e redistribuir suas melhorias para o público em geral.

|  |  |
| --- | --- |
|  |  |

E. ( X ) Liberdade de utilizar a propriedade intelectual alheia para qualquer fim.

**Questão 2:** Marque quais das afirmações abaixo seriam bons motivos para se utilizar o R para aplicações estatísticas:

A. ( ) Por default manipula facilmente grandes quantidades de dados, pois gerencia a memória dinamicamente, mantendo a maior parte dos dados no HD por default.

|  |  |
| --- | --- |
|  |  |

B. ( X ) Possui uma rede global colaborativa de desenvolvimento de pacotes e soluções diversas, fornecendo versatilidade na sua aplicação.

|  |  |
| --- | --- |
|  |  |

C. ( ) É de propriedade de uma sólida empresa que fornece suporte e garante a integridade e a qualidade das aplicações fornecidas.

|  |  |
| --- | --- |
|  |  |

D. ( X ) É licenciado pelo GNU – GLP, portanto seu uso é livre para qualquer fim e gratuito.

|  |  |
| --- | --- |
|  |  |

E. ( ) A empresa proprietária do software possui equipe extremamente especializada que sempre implementa o que há de mais moderno em *analytics* nas atualizações do software.

**Questão 3 – revisão em regressão logística:** Preocupada com o alto *turnover* de sua equipe, a HBAT realizou uma pesquisa de clima organizacional. Cada colaborador respondeu a um questionário com 21 perguntas escolhidas pela consultoria escolhida, as perguntas relacionadas abaixo.

|  |  |  |
| --- | --- | --- |
| **Item** | **Tipo de escala** | **Descrição** |
| JS1 | Likert 0 a 10 | Lenvando tudo em conta, sinto-me muito satisfeito quando penso em meu emprego |
| OC1 | Likert 0 a 10 | Meu trabalho na HBAT me dá uma sensação de realização |
| OC2 | Likert 0 a 10 | Estou disposto a fazer além do normalmente esperado, para ajudar a HBAT |
| EP1 | Likert 0 a 10 | Estou confortável com meu ambiente físico na HBAT |
| OC3 | Likert 0 a 10 | Sinto lealdade para com a HBAT |
| OC4 | Likert 0 a 10 | Sinto orgulho de dizer aos outros que trabalho para a HBAT |
| EP2 | Likert 0 a 10 | O local onde trabalho é planejado para me ajudar a fazer meu trabalho melhor |
| EP3 | Likert 0 a 10 | Há pouco sobstáculos que podem me tornar menos produtivo em meu local de trabalho |
| AC1 | Likert 0 a 10 | O quão feliz você está com o trabalho de seus colegas |
| EP4 | Dif. Semântico | Qual termo melhor descreve seu ambiente de trabalho na HBAT |
| JS2 | Dif. Semântico | Quando você pensa em seu trabalho, quão satisfeito você fica |
| JS3 | Dif. Semântico | O quão satisfeito você está com seu trabalho atual na HBAT |
| AC2 | Dif. Semântico | Como você se sente com relação a seus colegas |
| SI1 | Likert 5 p | Atualmente não estou procurando emprego |
| JS4 | Likert 5 p | Como empregado, o quão satisfeito você está com a HBAT |
| SI1 | Likert 5 p | Raramente olho a lista de empregos na internet |
| JS5 | Percentual | Indique o % de satisfação com seu emprego atual na HBAT |
| AC3 | Likert 5p | Com que frequencia você faz coisas com seus colegas de trabalho |
| SI3 | Dif. Semântico | Não tenho interesse em procurar emprego no próximo na |
| AC4 | Dif. Semântico | Geralmente o quanto seus colegas se parecem com você |
| SI4 | Likert 5p | Qual a possibilidade de você estar trabalhando na HBAT pelos próximos 12 meses |

A variável evasao é uma variável indicadora de se o colaborador deixou a empresa nos 6 meses seguintes à pesquisa.

A direção da empresa quer entender quais das perguntas respondidas melhor explicam o fenômeno da evasão afim de adotar um plano de ação.

Utilize o trecho de código abaixo para responder à questão:

hbat400<-read.csv(“hbat\_clima.csv”, header=T)

fit.min.hbat400<- glm(evasao~1,family = binomial(link = "logit"),data=hbat400)

stp.hbat400<-step(fit.min.hbat400,

direction="both",

scope=(~

JS1 + JS2 + JS3 + JS4 + JS5 +

OC1 + OC2 + OC3 + OC4 +

EP1 + EP2 + EP3 + EP4 +

AC1 + AC2 + AC3 + AC4 +

SI1 + SI2 + SI3 + SI4

))

summary(stp.hbat400)

Analisando a relação entre as perguntas e a evasão observada, marque todas as afirmativas corretas (mais de uma pode ser verdadeira):

1. ( X ) Dada a presença das variáveis SI4, SI1 e JS5, não há mais variáveis relevantes adotando-se 10% de significância.
2. ( ) Adotando-se 10% de significância, podemos afirmar que **evasão** é explicada **somente** pelas variáveis SI4, SI1 e JS5, não havendo mais variáveis relevantes a se investigar.
3. ( X ) O modelo contendo as variáveis SI4, SI1 e JS5 é um modelo a se considerar, no entanto não é necessariamente a melhor solução.
4. ( X ) SI3 não apresenta relação estatisticamente significante com evasão dadas as variáveis SI4, SI1 e JS5.
5. ( X ) SI3 pode ter relação com evasão se retirarmos do modelo uma ou mais das variáveis consideradas até o momento (SI1, SI4 e JS5).

Apesar da função step considerar para entrada/saída das variáveis no modelo o AIC (Akaike Information Criteria), alterou-se o parâmetro de k para adotar o nível de significância de 0,10. Assim, considerou-se verdadeiro o primeiro item, pois dado as variáveis que estão presentes no modelo, nenhuma outra apresentou p-valor inferior a 0,10. Em outras palavras, nenhuma variável adicionou mais informação relevante ao modelo ao ponto dessa contribuição compensar a perda de grau de liberdade.

**Questão 4:** O analista de RH da HBAT, preocupado com algumas variáveis não terem sido selecionadas pelo método automático indicado na questão 2, executou o trecho de código abaixo.

summary(glm(evasao~SI1,family = binomial(link = "logit"),data=hbat400))

summary(glm(evasao~SI2,family = binomial(link = "logit"),data=hbat400))

summary(glm(evasao~SI3,family = binomial(link = "logit"),data=hbat400))

summary(glm(evasao~SI4,family = binomial(link = "logit"),data=hbat400))

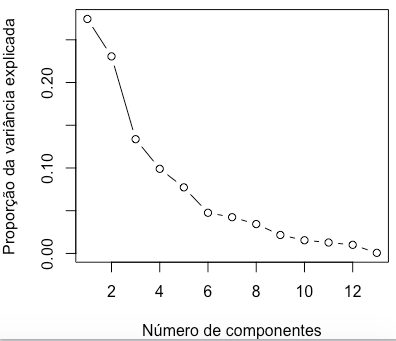
cor(hbat400[,c(16,18,21)])

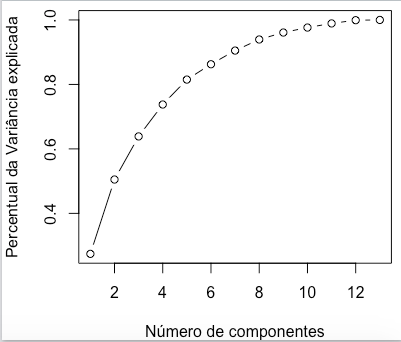
Assinale as conclusões que os resultados indicam (mais de uma pode ser verdadeira).

1. ( ) A variável SI3 não apresenta relação com a evasão observada, por isso não apareceu no resultado do código mostrado na questão 2.
2. ( X ) Todas as variáveis investigadas (SI1, SI2, SI3 e SI4) apresentam relação com Evasão se testadas individualmente, no entanto estão correlacionadas entre si, por isso não aparecem todas significantes ao mesmo tempo.
3. ( X ) Observando que as variáveis SI2 e SI3 apresentam significância individualmente e saem na presença de outras variáveis, provavelmente existe um efeito de multicolinearidade que proporcionou este fenômeno.
4. ( ) As conclusões não são consistentes porque o código foi rodado em bases diferentes.

**Questão 5:** O analista de RH da HBAT diagnosticou as variáveis como multicolineares e tentou uma abordagem por Componentes Principais para resolver o problema.

Observando o gráfico de explicação da variância abaixo, adotando como critério que cada componente deve explicar pelo menos 5% de variância, o que devemos esperar da nova análise?





1. ( ) Ele vai escolher 8 componentes e obterá certamente um modelo mais interpretável.
2. ( X ) Ele vai escolher 5 componentes e deve resolver o problema de multicolinearidade, no entanto seu modelo sofrerá críticas devido ao uso de uma componente principal como entrada.
3. ( ) O modelo ficará certamente com pior ajuste pois como ele vai utilizar apenas 3 componentes vai considerar apenas cerca de 60% da variância total.
4. ( X ) O ajuste do modelo pode piorar tanto quanto melhorar.

**Questão 6 (desafio):** Estratégia para resolução de multicolinearidade com perda mínima de informação. Realize os passos abaixo para responder à questão que segue:

1. Construa uma análise de componentes principais nos dados da HBAT somente com as variáveis SI1, SI2, SI3 e SI4;
2. Construa uma variável com o escore da primeira componente – isto, é a projeção das 4 variáveis na primeira componente principal da análise feita em (1).
3. Faça um modelo logístico com esta variável.

Compare os indicadores de KS, GINI e teste de Hosmer & Lemeshow do modelo obtido com os do modelo anterior (com as variáveis SI1, SI4 e JS5).

O que podemos concluir? (mais de uma afirmação pode ser verdadeira)

1. ( X ) O modelo com a componente principal é pior em nível de discriminação, pois os indicadores de GINI e KS são piores;
2. ( ) O modelo com três variáveis tem melhor ajuste, pois a estatística do teste de Hosmer e Lemeshow está melhor.
3. ( X ) Os modelos são equivalentes em termos de discriminação e ajuste, pois possuem indicadores com valores muito semelhantes.
4. ( ) O modelo com a componente principal possui pior ajuste, pois a estatística do teste de Hosmer e Lemeshow é maior, e melhor discriminação, pois os indicadores de KS e GINI são maiores.

Em relação ao item “c” avalio que os dois modelos estão na mesma faixa de discriminação (boa discriminação, faixa de 0,30 < KS < 0,40). O modelo stepwise apresenta KS = 0,37 e Gini = 0,47, enquanto que o modelo utilizando componentes principais possui KS = 0,32 e Gini = 0,44. Já em relação ao ajuste do modelo, os dois aceitaram H0 no teste de Hosmer e Lemeshow.

**SCRIPT EM R**

##################

### DIRETORIO ###

##################

getwd()

setwd("/Volumes/KINGSTON/FGV/Clusters e matrizes/Aula 02")

#################

### LIBRARIES ###

#################

library(dplyr)

library(hmeasure)

library(ResourceSelection)

#################

#### DATASET ###

#################

hbat400<-read.csv("hbat\_clima.csv")

##################

### QUESTÃO 3 ###

##################

fit.min.hbat400<- glm(evasao~1,family = binomial(link = "logit"),

data=hbat400)

stp.hbat400<-step(fit.min.hbat400,

direction="both",

scope=(~

JS1 + JS2 + JS3 + JS4 + JS5 +

OC1 + OC2 + OC3 + OC4 +

EP1 + EP2 + EP3 + EP4 +

AC1 + AC2 + AC3 + AC4 +

SI1 + SI2 + SI3 + SI4

))

summary(stp.hbat400)

# A função step utiliza o AIC como critério para a entrada/saída das variáveis no modelo.

# Como default o p-valor é de 0.15 (0.1573).

# Para alterar para o p-valor para 0.10, basta modificar o parâmetro k para 2.7

# Fonte: https://stats.stackexchange.com/questions/97257/stepwise-regression-in-r-critical-p-value

stp.hbat400\_2<-step(fit.min.hbat400,

direction="both",

scope=(~

JS1 + JS2 + JS3 + JS4 + JS5 +

OC1 + OC2 + OC3 + OC4 +

EP1 + EP2 + EP3 + EP4 +

AC1 + AC2 + AC3 + AC4 +

SI1 + SI2 + SI3 + SI4

), k = 2.7)

summary(stp.hbat400\_2)

##################

### QUESTÃO 4 ###

##################

summary(glm(evasao~SI1,family = binomial(link = "logit"),data=hbat400))

summary(glm(evasao~SI2,family = binomial(link = "logit"),data=hbat400))

summary(glm(evasao~SI3,family = binomial(link = "logit"),data=hbat400))

summary(glm(evasao~SI4,family = binomial(link = "logit"),data=hbat400))

cor(hbat400[,c(16,18,21)])

##################

### QUESTÃO 6 ###

##################

dataset\_q6 <- select(hbat400, starts\_with("SI"))

pca\_q6 <- princomp(dataset\_q6, cor = F) # utilizei a matriz de covariância, pois os dados já estão na mesma escala (Likert)

pca\_q6

pca\_q6$loadings # carga fatorial

pca\_q6$sdev # desvio padrão das componentes

plot(pca\_q6, type="l", main = "Variância das Componentes Principais")

componente1 <- pca\_q6$scores[,1]

componente1

hbat\_q6 <- as.data.frame(cbind(evasao=hbat400$evasao,componente1))

# head(hbat\_q6)

tail(hbat\_q6)

fit\_q6<- glm(evasao~.,family = binomial(), # o link = logistic é default

data=hbat\_q6)

summary(fit\_q6)

# Avaliação do ajuste do modelo - Teste Hosmer-Lemeshow

# 1) Modelo de Regressão Logística com Stepwise

hoslem.test(stp.hbat400$y, fitted(stp.hbat400))

# 2) Modelo de Regressão Logística com Componente Principal (SI)

hoslem.test(fit\_q6$y, fitted(fit\_q6))

# Avaliação do grau de discriminação do modelo

# 1) Modelo de Regressão Logística com Stepwise

pred\_evasao\_stp <- NULL

pred\_evasao\_stp <- predict(stp.hbat400,newdata = pred\_evasao\_stp, type = "response")

HMeasure(hbat400$evasao, pred\_evasao\_stp)$metrics

# 2) Modelo de Regressão Logística com Componente Principal (SI)

pred\_evasao\_pca <- NULL

pred\_evasao\_pca <- predict(fit\_q6,newdata = pred\_evasao\_pca, type = "response")

HMeasure(hbat400$evasao, pred\_evasao\_pca)$metrics