**NOMBRE:**

**Parte 1: Conceptos Modelamiento**

**Parte 2. Python**

**Parte 3. R y SAS**

**Parte 4. Caso**

Tip. Sea conciso y claro en sus argumentos.

**Parte 1: Conceptos Modelamiento**

**TEORÍA**

1. Explique las etapas que consideraría para el desarrollo de un modelo estadístico para originación.
2. Con la matriz de confusión presentada a continuación, calcule y explique las métricas presentadas.

|  |  |  |
| --- | --- | --- |
| n= 192 | Predicted: 0 | Predicted: 1 |
| Actual: 0 | 118 | 12 |
| Actual: 1 | 47 | 15 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Métrica** | **Fórmula** | **Resultado** | **Explicación** |
| Recall | True Positives /(True Positives +False Negatives) |  |  |
| Accuracy | (True Positives +True Negatives )/n |  |  |
| Misclassification Rate | (False Positives +False Negatives )/n |  |  |
| Specificity | True Negatives /(True Negatives +False Positives) |  |  |

1. En problemas de clasificación, un set de datos desbalanceado puede causar problemas a múltiples algoritmos de aprendizaje. Estos problemas se caracterizan por la proporción desigual de casos disponibles para cada clase del problema. Una forma de abordar este problema es volver a muestrear el conjunto de datos para compensar este desequilibrio con la esperanza de llegar a un límite de decisión más sólido y justo de lo que lo haría de otra manera. Las técnicas de remuestreo se dividen en dos categorías. Explíquelas, mencione una ventaja y una desventaja:

|  |  |  |  |
| --- | --- | --- | --- |
| **Método** | **Explicación** | **Ventaja** | **Desventaja** |
| Undersampling |  |  |  |
| Oversampling |  |  |  |

1. Mencione 4 algoritmos de Machine Learning, explique en que consiste la metodología del algoritmo y mencione ejemplos de posibles usos.

**Parte 2 Python**.

Para los siguientes códigos, poner el resultado que se obtiene al ejecutarlo:

1.

p = lambda q: q \* 2

m = lambda q: q \* 3

y = 2

y = p(y)

y = m(y)

y = p(y)

print y

Respuesta:

2. nameList = ['Hunter', 'Punirk', 'Bonl', 'Dino']

  print nameList[1][-1]

Respuesta:

3.

list = [1, 80, 6, None, (1, 4, 3, 4, 7), ['Peeks', 'Fin', ‘Toner’]]

print len(list)

Respuesta:

4.

list1 = ['hola', 'Colombia', 1998, 2005]

list2 = [1, 2, 3, 4, 5, 6, 7 ]

print "list1[0]: ", list1[0]        #statement 1

print "list1[0]: ", list1[-2]       #statement 2

print "list1[-2]: ", list1[1:]      #statement 3

print "list2[1:5]: ", list2[1:5]    #statement 4

Respuesta:

5. Menciones dos paquetes de Python y explique sus principales usos.

**Parte 3. R y SAS**

1. R es un entorno y lenguaje de programación con un enfoque al análisis estadístico.

 Verdadero  Falso

1. ¿Cuál sería la instrucción para obtener ayuda acerca de la función Max?

 max.help()  max.help  help(max)  help.max

1. ¿Cuál es el resultado de la siguiente expresión?

> 2+3+4+10

 18  19  20  Nos muestra un mensaje de error

1. ¿Cuál es el resultado de la siguiente expresión?

> 2 \* 2,5

 5  4,5  4   Error: unexpected ',' in "2 \* 2,"

1. ¿Qué signo(s) se usa para escribir comentarios?

  /\* y \*/  //  #   $

1. ¿Qué operador se utiliza para obtener el cociente entero?

 %  %/  %/%  %%

1. ¿Cuál es el resultado de la siguiente expresión?

> 9%%2

 4.5  4,5  4  1

1. ¿Cuál es el resultado de la siguiente expresión?

> sqrt(9)

 2  3   4   9

9. ¿Qué expresión me devuelve el siguiente resultado?

3.1416

pi  print(pi,5)  print(pi,4)  print(pi,3)

10. Para crear una variable y que contenga el valor de pi multiplicado por 3. Utilizaría la siguiente expresión:

y=pi\*3

 Verdadero  Falso

11. El siguiente código crea una lista de objetos.

lst <- list(numbers = 1:10, letters = letters, boolean = c(TRUE, FALSE))

lst

*## $numbers*

*## [1] 1 2 3 4 5 6 7 8 9 10*

*## $letters*

*## [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j" "k" "l" "m"*

*## [14] "n" "o" "p" "q" "r" "s" "t" "u" "v" "w" "x" "y" "z"*

*## $boolean*

*## [1] TRUE FALSE*

Que retorna?

lst$numbers

lst[1]

lst[[1]]

12. El siguiente código crea una función que retorna una lista. Asuma que corre el código:

make\_list <- function() {

list(date = Sys.Date(),

time = Sys.time(),

timezone = Sys.timezone())

}

*## $date*

*## [1] "2015-03-12"*

*##*

*## $time*

*## [1] "2015-03-12 16:58:13 EDT"*

*##*

*## $timezone*

*## [1] "America/New\_York"*

Cuál la seria la sintaxis correcta para imprimir el ítem “time”:

make\_list(time)  make\_list() make\_list.time make\_list()$time

13. Cuál es la diferencia entre una matriz y un dataframe:

14. The following SAS program is submitted:

***data work.total;***

***set work.salary(keep = department wagerate);***

***by department;***

***if first.department then payroll = 0;***

***payroll + wagerate;***

***if last.department;***

***run;***

The SAS data set named WORK.SALARY contains 10 observations for each department, currently ordered by DEPARTMENT.

Which one of the following is true regarding the program above?

A. The BY statement in the DATA step causes a syntax error.

B. FIRST.DEPARTMENT & LAST.DEPARTMENT are variables in WORK.TOTAL dataset.

C. The values of the variable PAYROLL represent the total for each department in the WORK.SALARY data set.

D. The values of the variable PAYROLL represent a total for all values of WAGERATE in the WORK.SALARY data set.

15. The following SAS program is submitted:

***data test;***

***set sasuser.employees;***

***if 2 le years\_service le 10 then***

***amount = 1000;***

***else if years\_service gt 10 then***

***amount = 2000;***

***else***

***amount = 0;***

***amount\_per\_year = years\_service / amount;***

***run;***

Which one of the following values does the variable AMOUNT\_PER\_YEAR contain if an employee has been with the company for one year?

A. 0

B. 1000

C. 2000

D. (missing numeric value)

**Parte 4. Caso**

La Gerencia de Ciencia de Datos es la encargada de desarrollar herramientas analíticas para la originación, el mantenimiento y la cobranza de los clientes del banco. Adicionalmente cuenta con información interna del cliente que es solicitada en el formulario de vinculación, así como también de información del buró de crédito acerca de su comportamiento en el mercado. Usted ha sido seleccionado como el nuevo Analista de Ciencia de Datos y le han asignado liderar todo el análisis de un portafolio de tarjeta de crédito marca compartida con un retail comercial.

El banco debe tomar la decisión de usar la propuesta generada por la Gerencia de Ciencia de Datos a la que usted pertenece o tercerizar el desarrollo y estudio con un outsourcing como Datacrédito, Cifin o Lisim. Usted es nuevo en la gerencia y le entregan tres propuestas de modelos de scoring de originación que fueron desarrollados por otra persona (archivo adjunto de Excel).

Dada la premura del proyecto y la decisión que debe tomar la alta gerencia, su misión es escoger el mejor modelo y presentarlo ante un comité conformado por el Gerente Comercial, el Gerente de Mercado, el Gerente de Riesgo y otras personas involucradas con el producto. El comité espera que usted argumente su decisión. ¡Buena suerte!

**Métricas de desempeño de los Modelos:**



