POTENCIALES CLIENTES BANCARIOS

Nombre: María Paz Santibáñez

Fecha: 24 noviembre 2023

OBJETIVO

El objetivo de nuestro trabajo es evaluar el rendimiento de modelos de aprendizaje supervisado y no supervisado para resolver un problema de clasificación binaria, y un problema de clustering respectivamente, utilizando el software R Studio y un conjunto de datos de una campaña de marketing telefónica de instituciones bancarias portuguesas realizadas entre mayo del 2008 hasta noviembre del 2010. Por un lado, se emplean modelos de regresión lineal, regresión ridge, regresión lasso, regresión logística y árboles de decisión para predecir si un cliente de una institución bancaria accederá a un deposito a plazo, y luego, se emplea un método de clustering K-means para segmentar a los clientes en clústeres según sus similitudes y patrones.

METODOLOGIA

El conjunto de datos incluye variables categóricas y numéricas que incluyen características personales de los clientes y detalles del contacto realizado antes y durante la campaña. Además, considera índices del mercado de ese periodo. Las imágenes 1 y 2 del Anexo, pág., 9 muestran el nombre y descripción de cada variable. Se descarta la variable "*duration*" ya que este atributo afecta en gran medida el objetivo de salida (por ejemplo, si duration = 0 entonces y = 'no'). Luego, se realiza un análisis de la calidad de los datos, en este se obtiene el porcentaje de datos nulos de todas las variables, se identifican los outliers mediante resúmenes, histogramas y box plot de las variables numéricas, y diagramas de Pareto en el caso las variables categóricas, y se imputan mediante la eliminación de las observaciones, se transforman las variables originales (carácter, numérica y entero) a variables tipo factor y numérica, donde estas últimas se estandarizan. Además, se obtiene la matriz de correlación que permitirá mejorar el modelo de regresión logística al eliminar la multicolinealidad. Para el modelo de regresión lineal, ridge y lasso se transforman todas las variables categóricas a numéricas mediante la codificación one-hot. Para el modelo de regresión logística se eliminan las variables numéricas que presentan multicolinealidad. Luego, se divide la data en 80% en muestra de entrenamiento y 20% muestra de validación, esta división permite entrenar un modelo en un conjunto de datos y luego probar su rendimiento en datos no vistos. Esto ayuda a evaluar cómo de bien el modelo puede generalizar a nuevos datos y a evitar el sobreajuste, y, por último, se utiliza la función *SMOTE* del paquete *DMwR* para balancear las clases de la variable dependiente una vez aplicada la codificación one-hot y *ovun.sample* del paquete *ROSE* para realizar un balance de las clases de la variable dependiente para el resto de los modelos, estas funciones combinan los métodos de sobremuestreo y submuestreo, es decir, crea nuevos ejemplos sintéticos en la clase minoritaria y elimina aleatoriamente algunas observaciones de la clase mayoritaria. Estas técnicas pueden ayudar a equilibrar las clases y a evitar el sesgo. La función *train* del paquete "caret" permite modelar una regresión lineal

usando el argumento method = lm y especificando que se utiliza la muestra de entrenamiento, además, junto con la función *trainControl* permite realizar una validación cruzada k-folds de 10 capas. Se evalúa el rendimiento del modelo para predecir las observaciones de la muestra de entrenamiento calculando el error cuadrático medio (MSE). Para la regresión ridge y lasso, se utiliza el paquete *glmnet* y se crean las matrices de entrenamiento y validación de las variables independientes, y el vector de respuesta de entrenamiento y validación. La funcion *glmnet* permite crear un modelo de regresión rigde definiendo el argumento alpha = 0, y un modelo de regresión lasso usando alpha = 1. La funcion *cv.glmnet* permite hacer una validación cruzada y obtener el valor del parámetro lambda que mejora que mejora la predicción para ambos modelos. Finalmente, utilizando el mejor valor del parámetro lambda se realizan las predicciones con la muestra de validación, y se obtiene el error cuadrático medio (MSE). Para la regresión logística se utiliza la funcion *glm* y la librería *caret* se establece el número de capas para la validación cruzada, este método permite evaluar el error de clasificación de los k-1 modelos y obtener la probabilidad de corte o el umbral que se utiliza para decidir entre las clases, que proporciona el mejor equilibrio entre la precisión y recall. A continuación, para modelar el árbol de decisión se utiliza la librería *tree* y la funcion *tree* se usan las observaciones de la muestra de entrenamiento, la funcion *cv.tree* permite realizar validación cruzada k-folds de 5 capas y obtener el tamaño optimo del árbol para realizar la poda. Por último, se utiliza la librería *clustMixtype* y la funcion *kproto* la cual permitirá agrupar en dos cluster con diferentes características.

RESULTADOS

El modelo de regresión lineal no es adecuado para resolver este problema puesto que la variable dependiente predicha se interpreta como continua, y en este caso nuestra variable es factor. Incluso aplicando codificación one-hot no se puede ejecutar el modelo y entrega el mensaje "*Error: wrong model type for classification*". Por otro lado, los modelos de regresión rigde y lasso pueden adecuarse para resolver problemas de clasificación binaria. El valor de beta que mejora la predicción es 0.000126 con un error cuadrático medio de 3.533 para regresión lasso, y un valor de beta igual a 0.0237855 con un error cuadrático medio de 3.318 para regresión rigde. La regresión logística es un modelo adecuado para problemas de clasificación donde la métrica F1 Score tiene un valor máximo de 0.72 con una probabilidad de corte de 0.4. El modelo de árbol de decisión es adecuado para resolver el problema de clasificación siendo 0.82 el accuracy de la predicción, y el tamaño optimo del árbol para podarlo es 3. Finalmente, el algoritmo k prototype entrega una buena segmentación de los dos grupos de clientes que quieren acceder a un deposito a plazo, ya que permite el uso de datos mixtos y además la funcion *clprofiles* permite observar visualmente las características de ambos grupos mediante boxplot de los dos grupos para las variables numéricas, y gráficos de barra para las variables categóricas.

CONCLUSIÓN

Los resultados demuestran que los modelos de regresión logística y árbol de decisión demostraron ser más efectivos para este problema de clasificación binaria. Estos modelos son capaces de capturar relaciones no lineales y complejas entre las características y la variable objetivo, lo que los hace más adecuados para problemas de clasificación.

REFERENCIAS

CRAN - Paquete clustMixType (r-project.org). (s.f.).

DMwR Package. (s.f.). Obtenido de https://cran.rproject.org/web/packages/DMwR/index.html

ROSE Package. (s.f.). Obtenido de https://cran.rproject.org/web/packages/ROSE/index.html

ANEXO

Imagen 1: Descripción variables numéricas

Variables numéricas	
nombre	descripción
age	edad
campaign	número de contactos realizados durante esta campaña y para este cliente
previous	número de contactos realizados antes de esta campaña y para este cliente
emp.var.rate	tasa de variación del empleo - indicador trimestral
cons.price.idx	índice de precios al consumo - indicador mensual
cons.conf.idx	índice de confianza del consumidor - indicador mensual
euribor3m	tipo de cambio euribor a 3 meses - indicador diario
nr.employed	número de empleados - indicador trimestral

Imagen 2: Descripción variables categóricas

Variables categóricas	
nombre	descripción
job	tipo de trabajo (categórico: 'administrador', 'obrero', 'emprendedor', 'criada', 'administración', 'jubilado', 'autónomo', 'servicios', 'estudiante', 'técnico', 'desempleado', 'desconocido')
marital	estado civil (categórico: 'divorciado', 'casado', 'soltero', 'desconocido')
education	educación (categórica: 'básico.4 años', 'básico.6 años', 'básico.9 años', 'bachillerato', 'analfabetos', 'curso.profesional', 'título universitario', 'desconocido')
default	¿tiene crédito en incumplimiento? (categórico: 'no', 'sí', 'desconocido')
housing	¿tiene préstamo para vivienda? (categórico: 'no', 'sí', 'desconocido')
loan	¿tiene préstamo personal? (categórico: 'no', 'sí', 'desconocido')
contact	tipo de comunicación del contacto (categórico: 'celular','teléfono')
month	último mes del año de contacto (categórico: 'ene', 'feb', 'mar', …, 'nov', 'dic ')
day_of_week	último día de contacto de la semana (categórico: 'lunes', 'martes', 'miércoles', 'jueves', 'viernes')
pdays	número de días que pasaron después de que el cliente fue contactado por última vez desde una campaña anterior (numérico; 999 significa que el cliente no contactado anteriormente)
poutcome	resultado de la campaña de marketing anterior (categórico: 'fracaso', 'inexistente', 'éxito')
y	¿el cliente ha suscrito un depósito a plazo? (binario: 'sí', 'no')

Imagen 3: Matriz de correlación variables numéricas

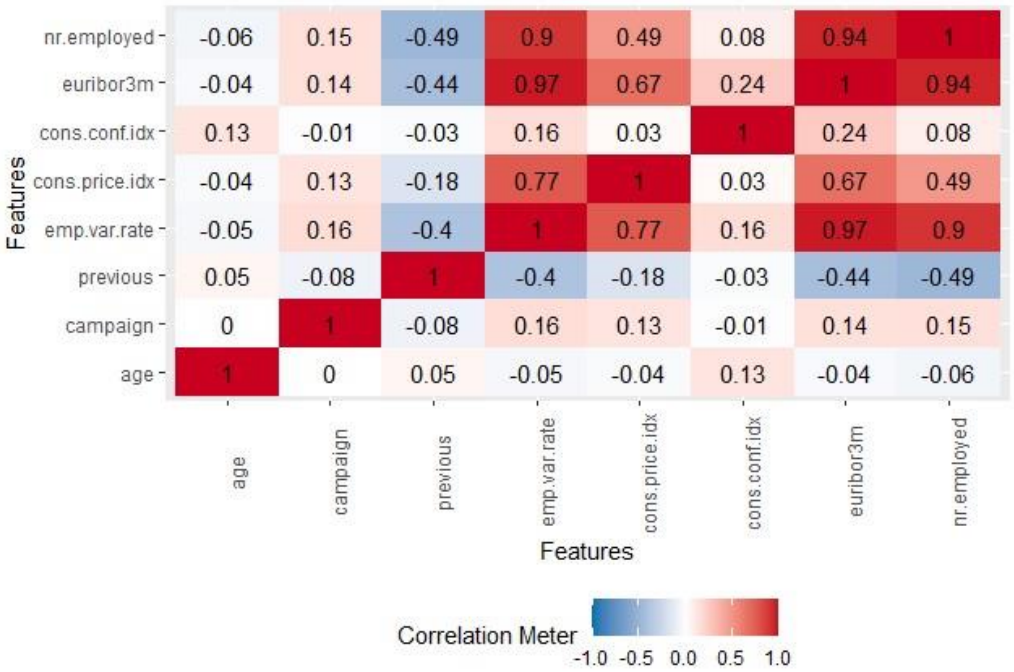


Imagen 4: Histogramas variables numéricas

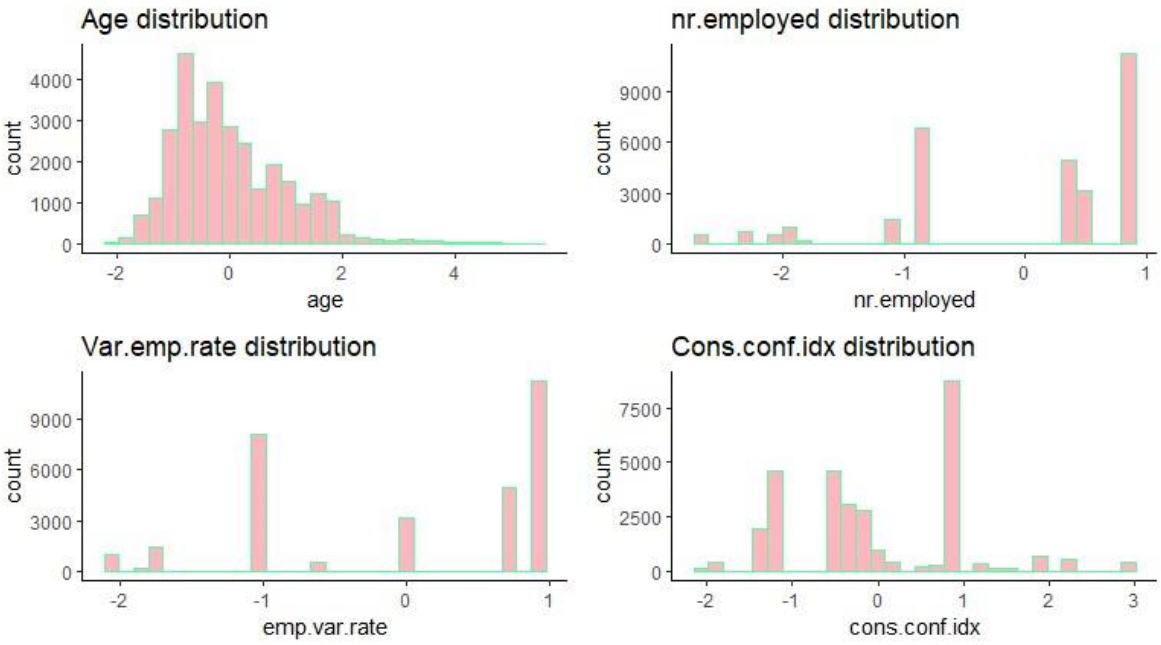


Imagen 5: Histogramas variables numéricas

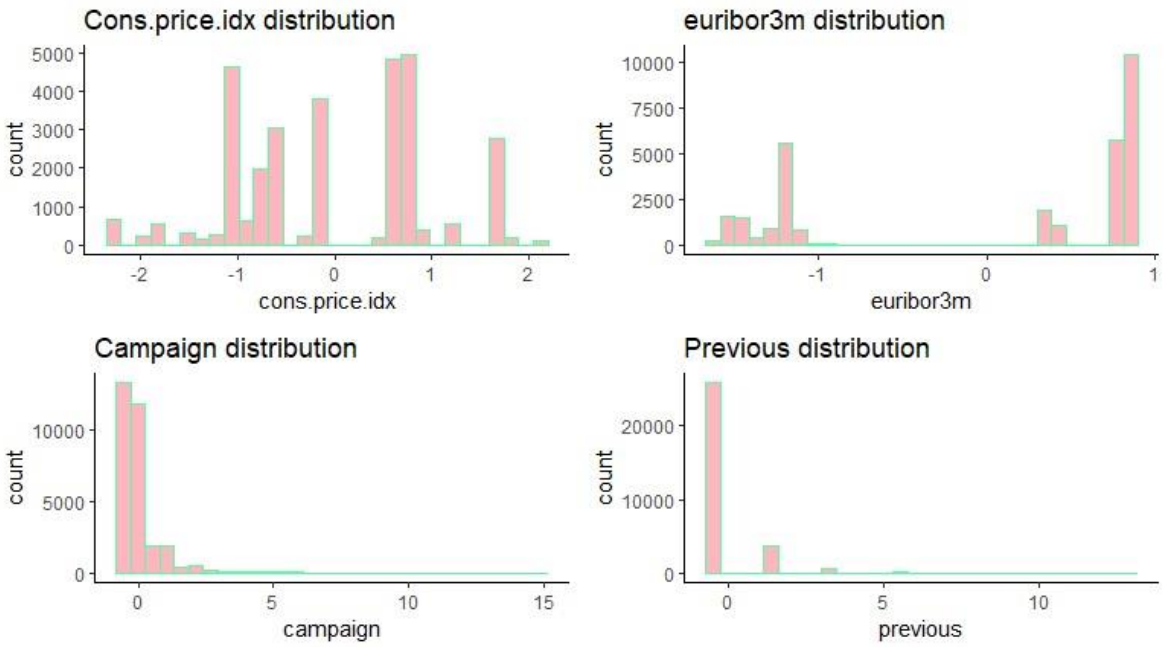
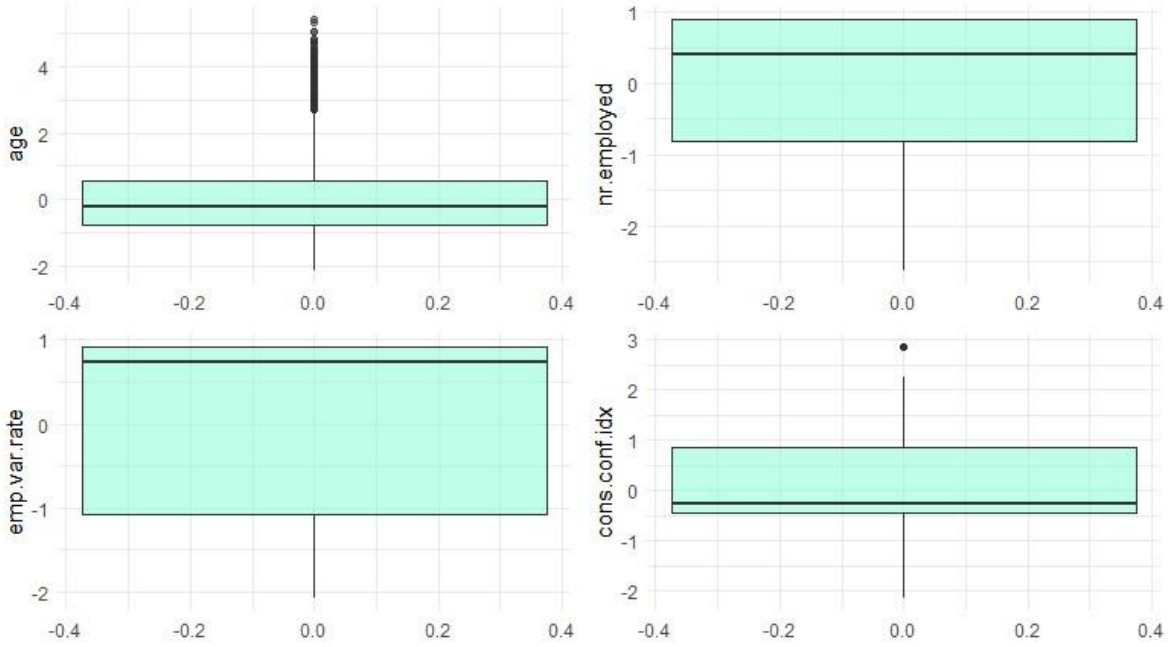


Imagen 6: Boxplot variables numéricas



Imagen

7: Boxplot variables numéricas

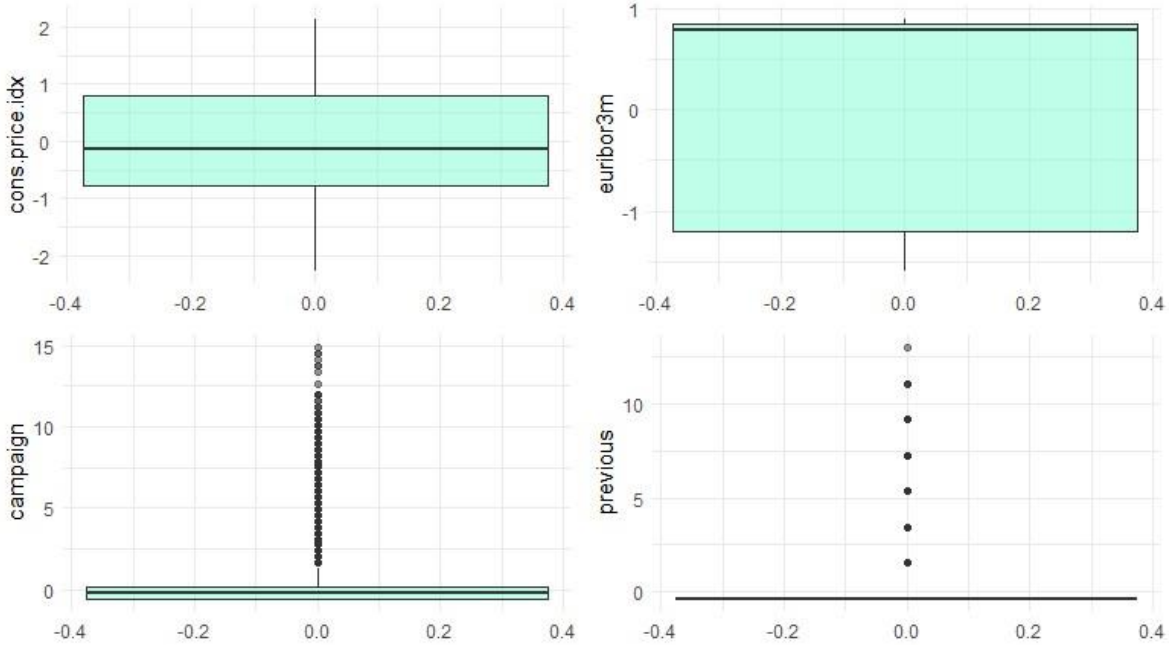
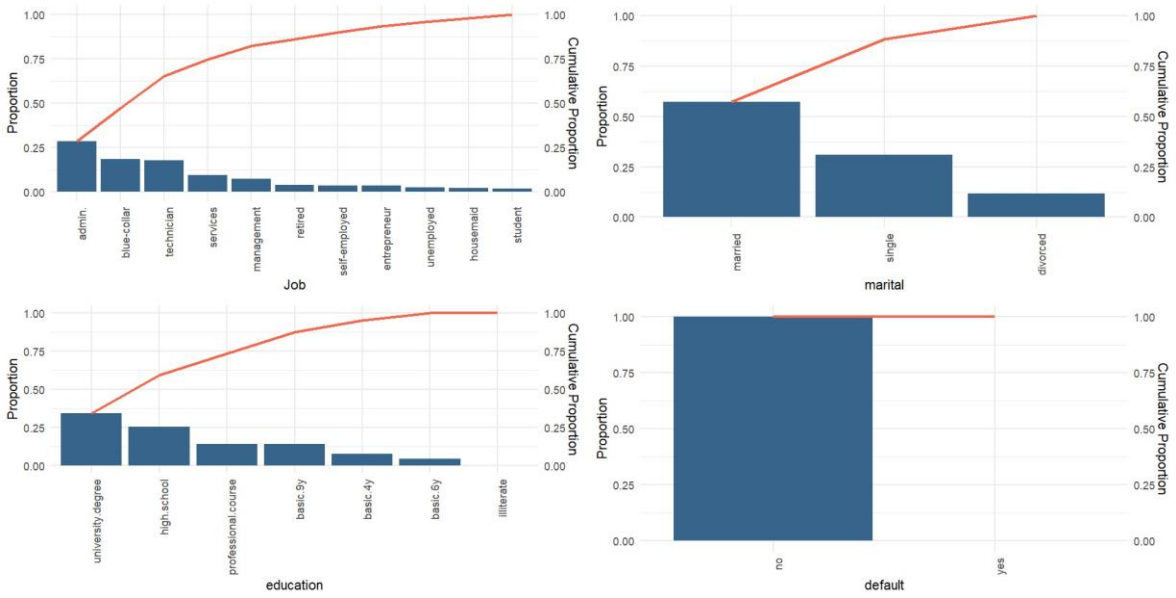


Imagen 8: Gráfico de Pareto variables categóricas



Imagen

9: Gráfico de Pareto variables categóricas

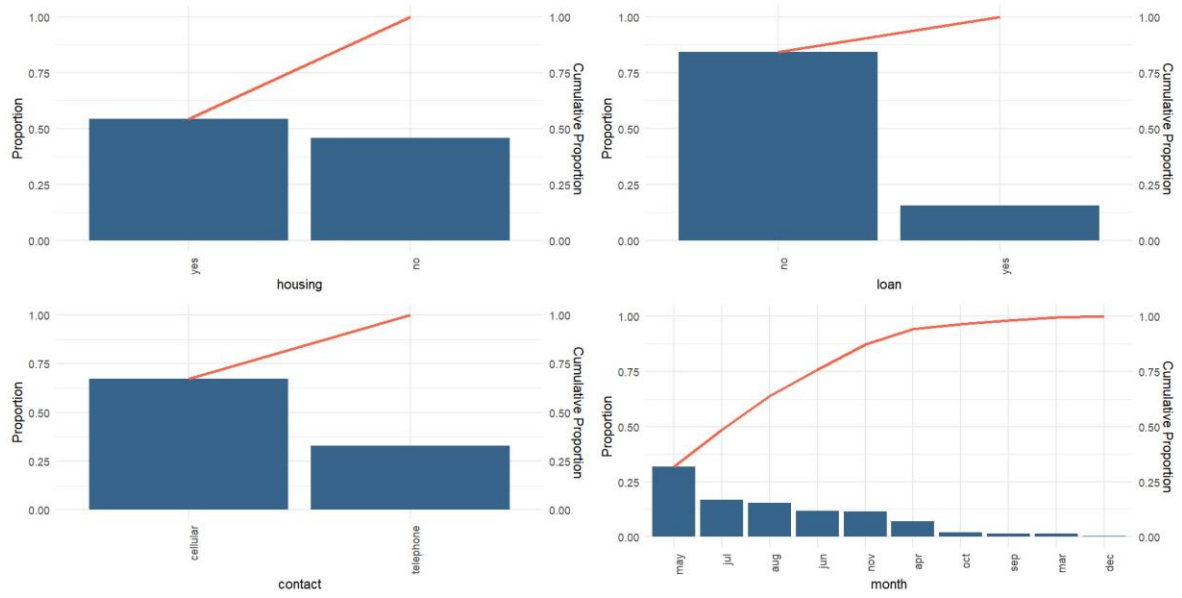
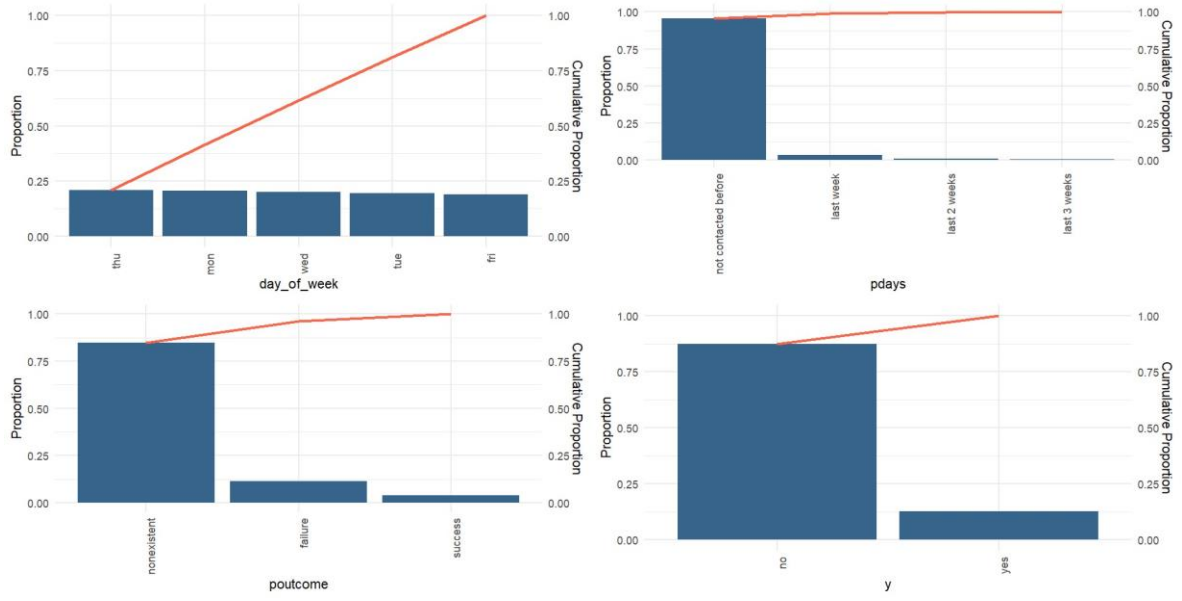
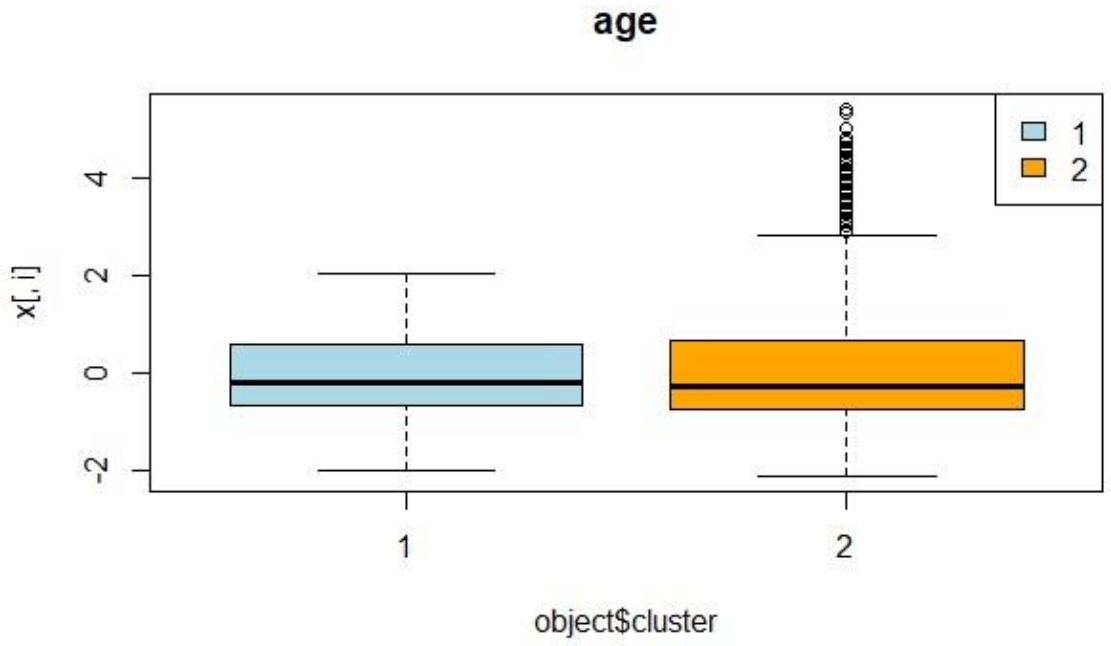


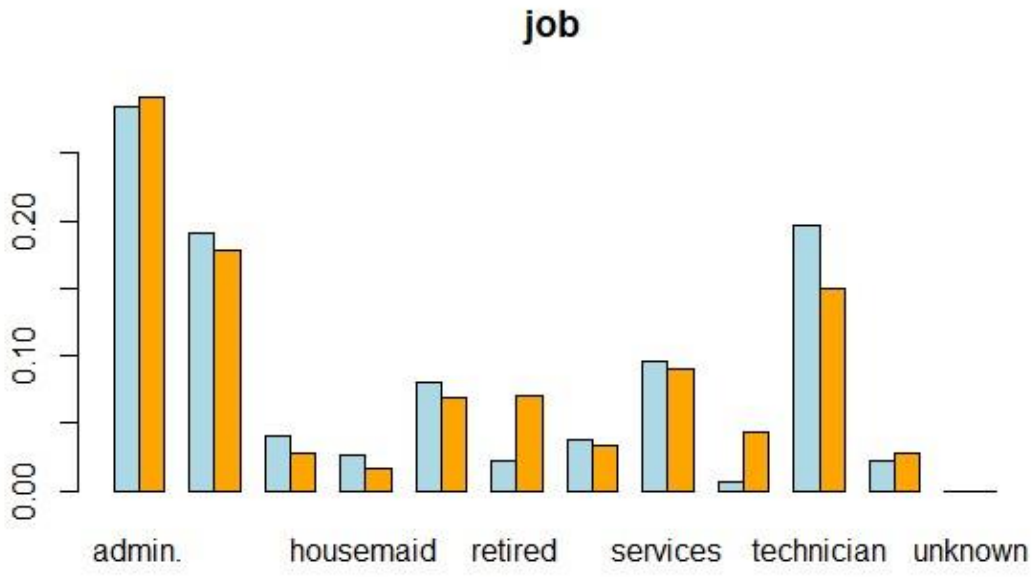
Imagen 10: Gráfico de Pareto variables categóricas

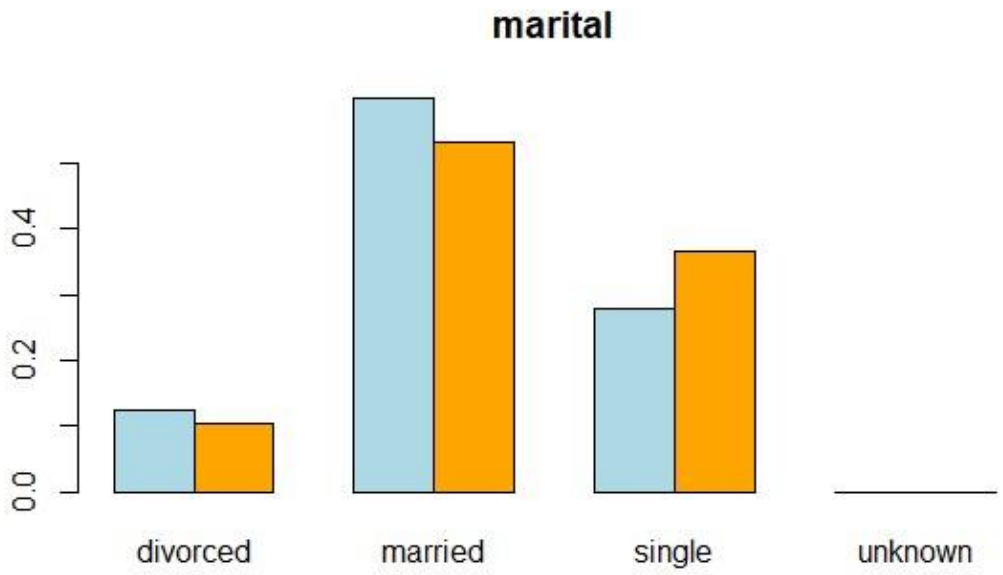


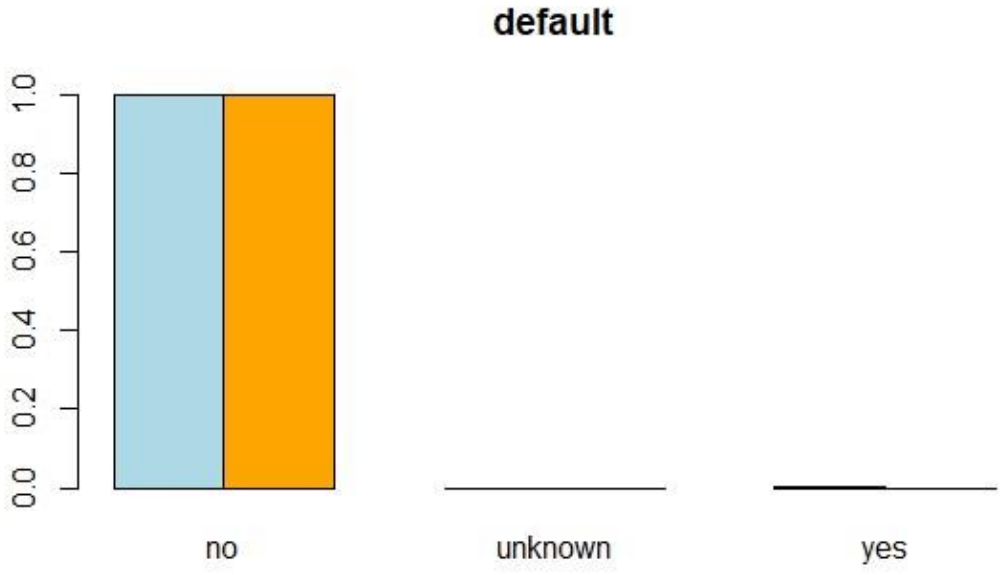
Imagen

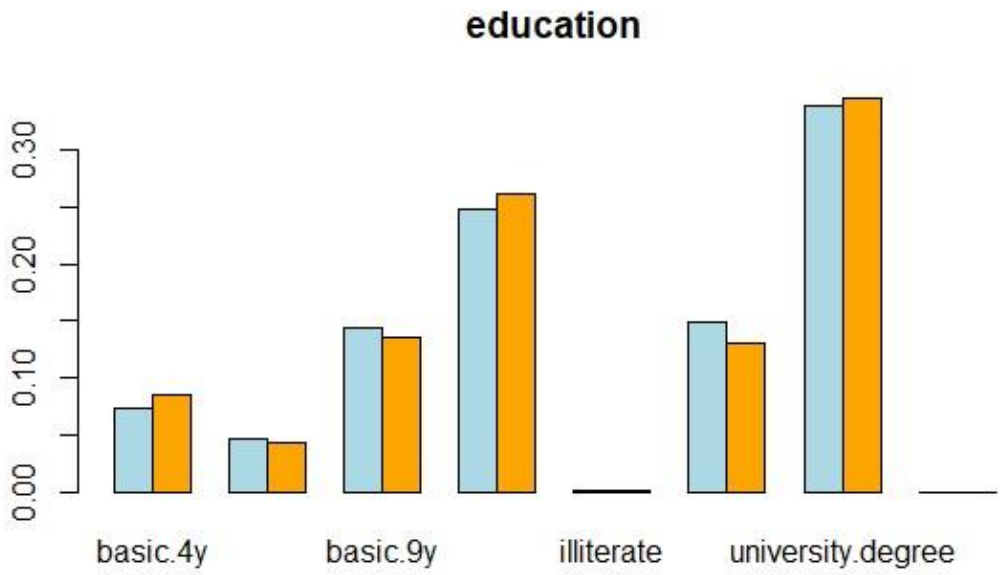
Imagen 11 a 29: Segmentación de clientes

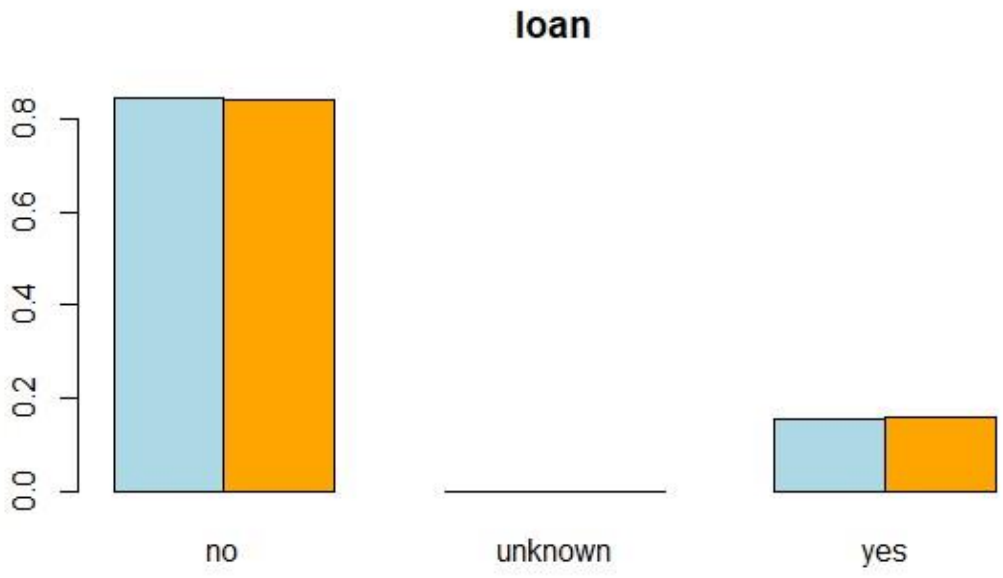


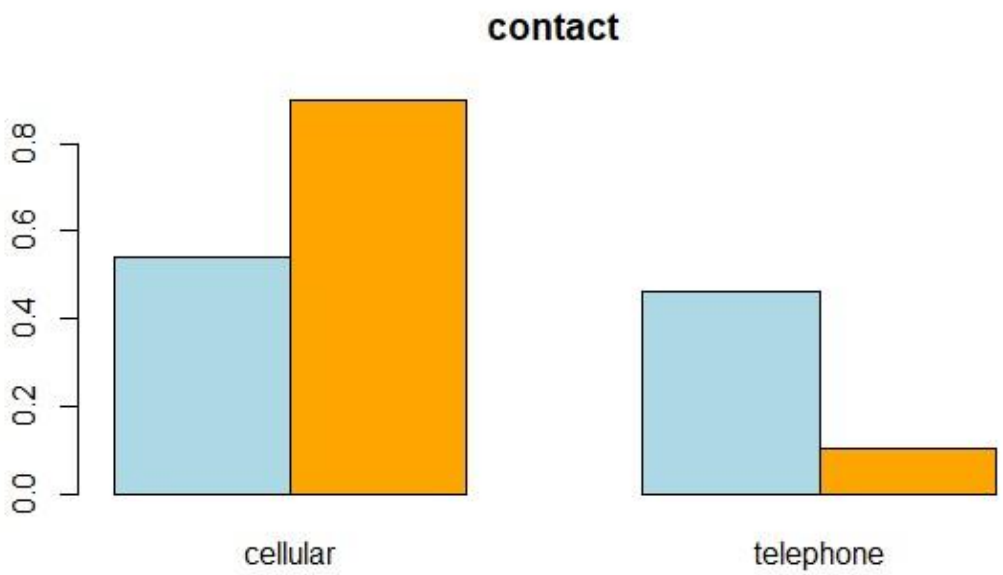


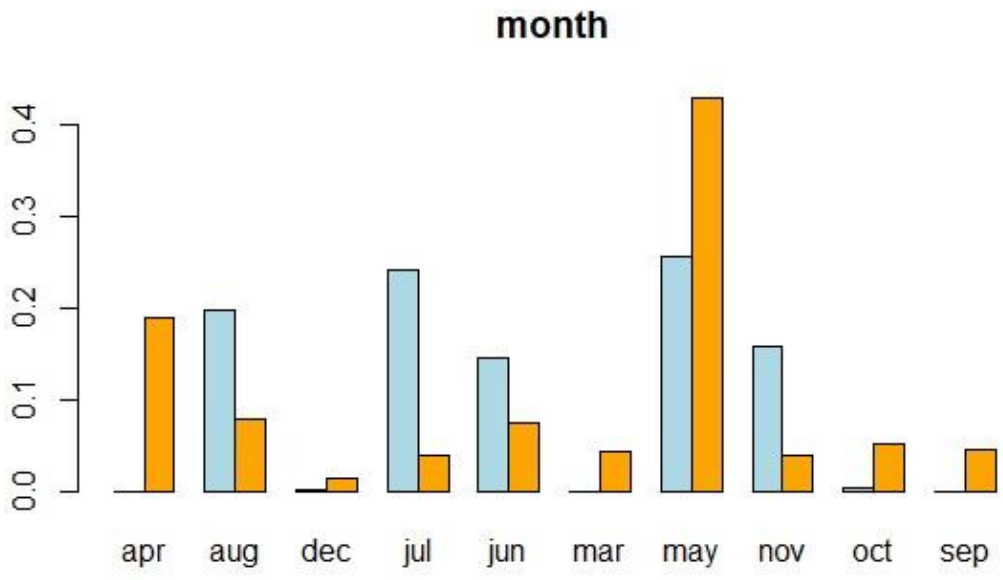

marital
0.0
0.2
0.4
divorced
married
single
unknown

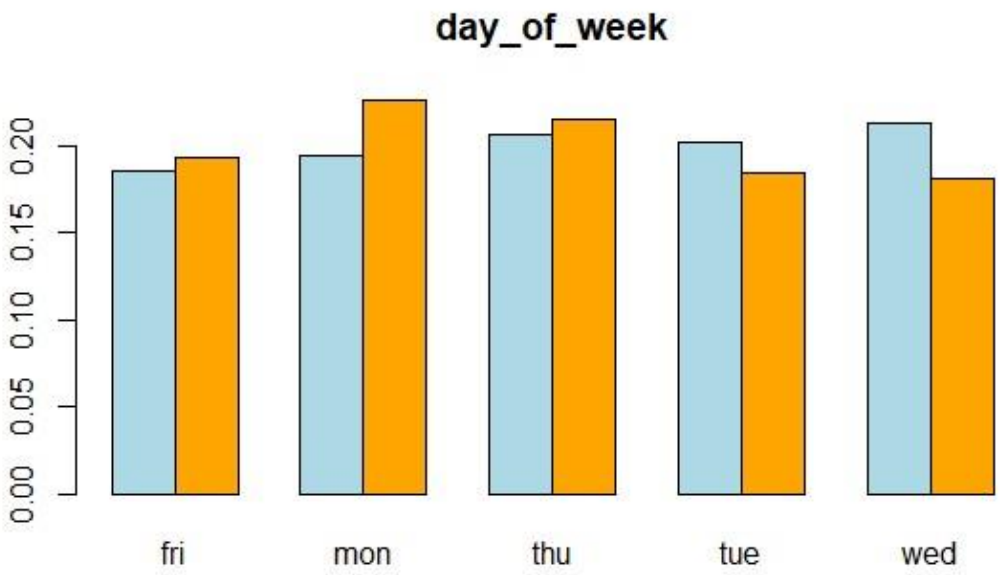

default
0.0
0.2
0.4
0.6
0.8
1.0
no
unknown
yes

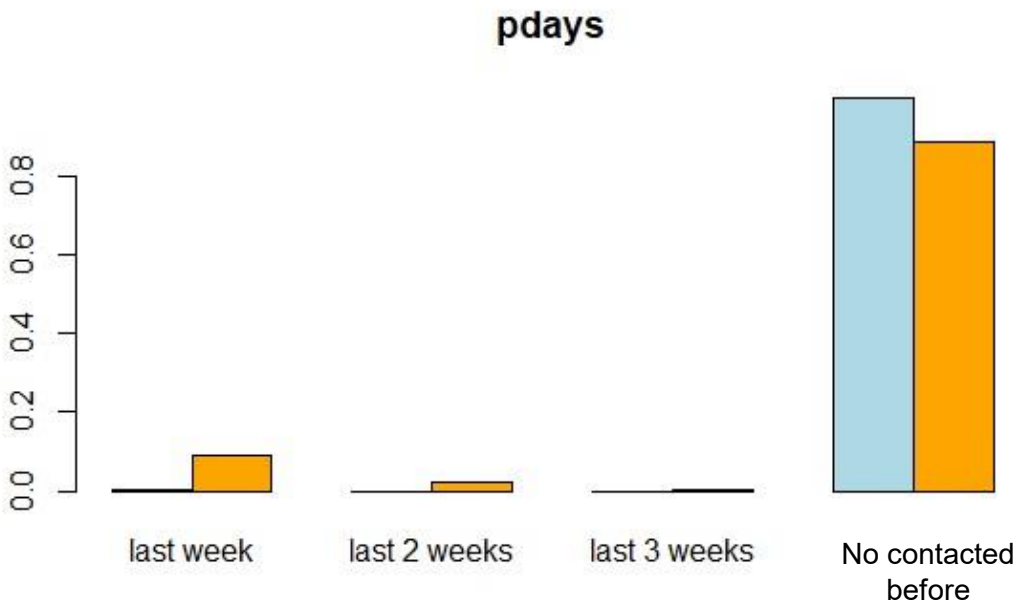

education
0.00
0.10
0.20
0.30
basic.4y
basic.9y
illiterate
university.degree

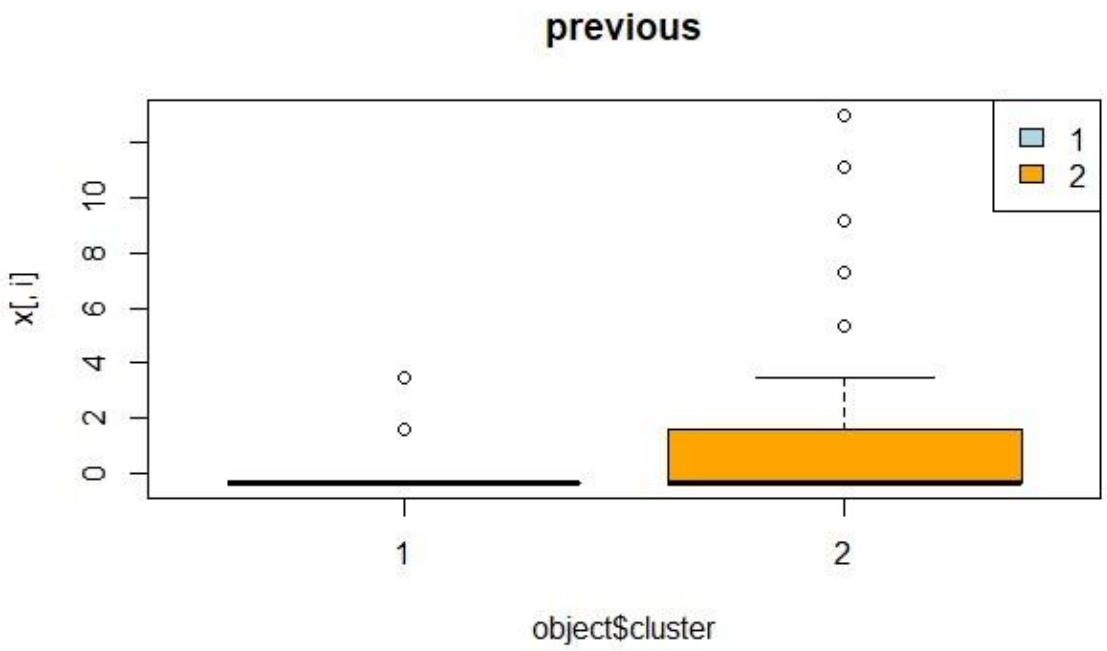

loan
0.0
0.2
0.4
0.6
0.8
no
unknown
yes

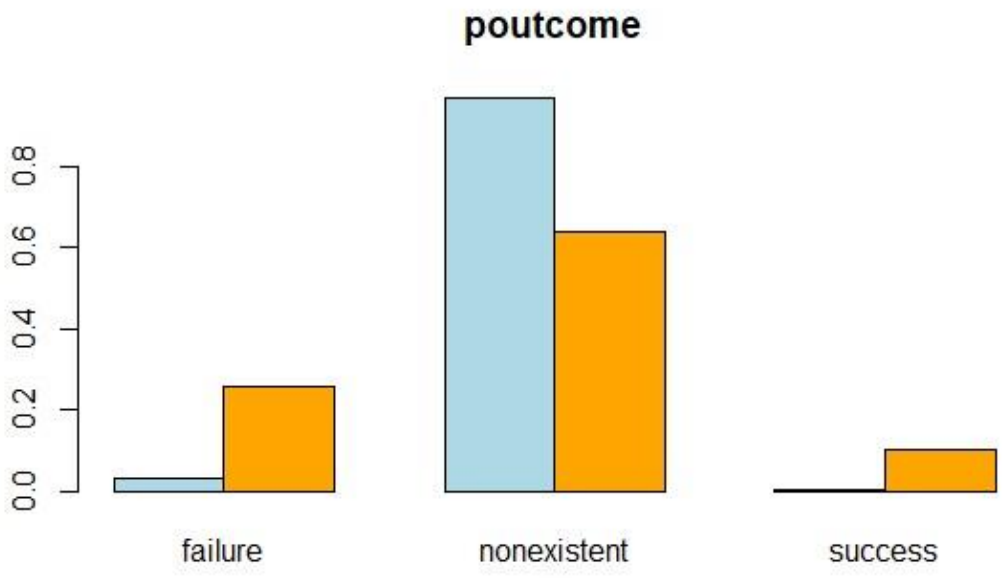

contact
0.0
0.2
0.4
0.6
0.8
cellular
telephone

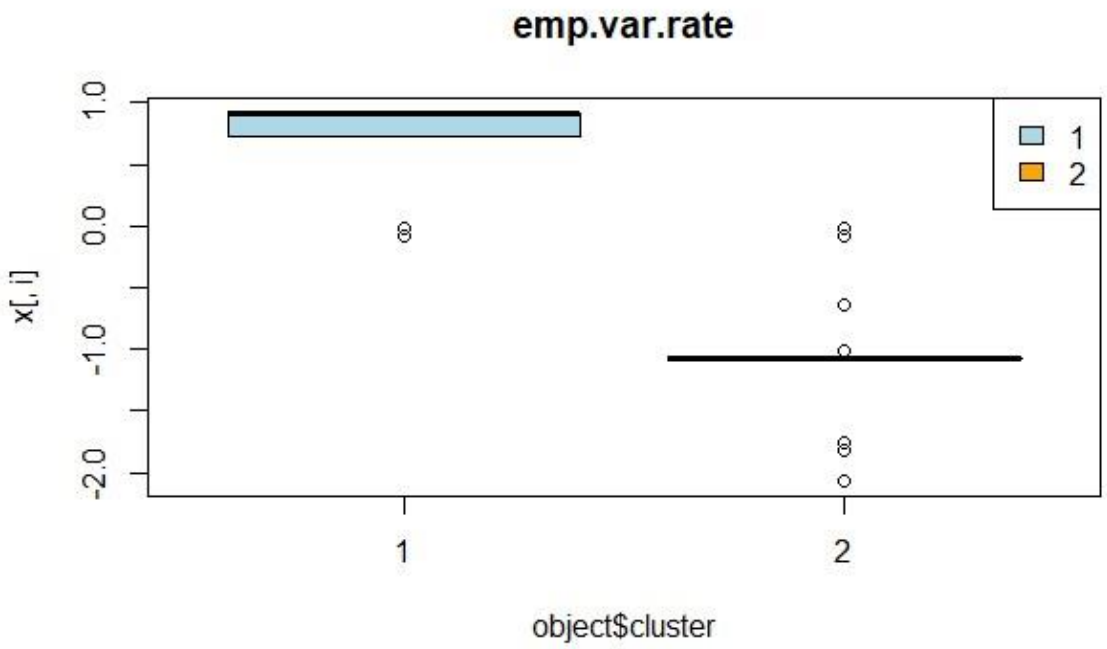

month
0.0
0.1
0.2
0.3
0.4
apr
aug
dec
jul
jun
mar
may
nov
oct
sep

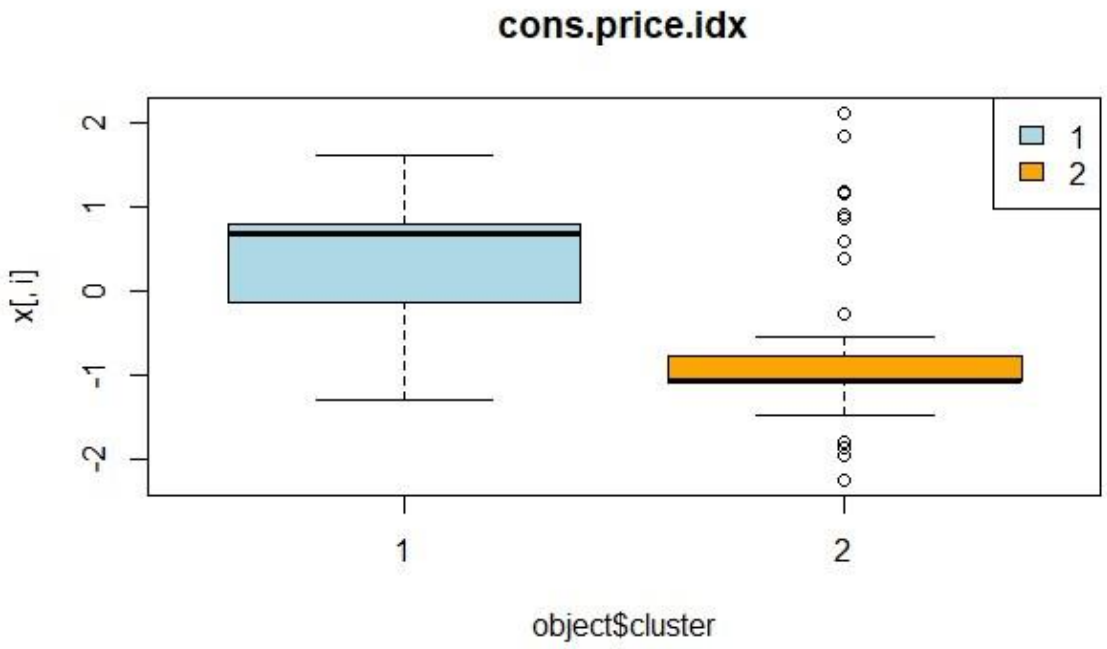

day_of_week
0.00
0.05
0.10
0.15
0.20
fri
mon
thu
tue
wed

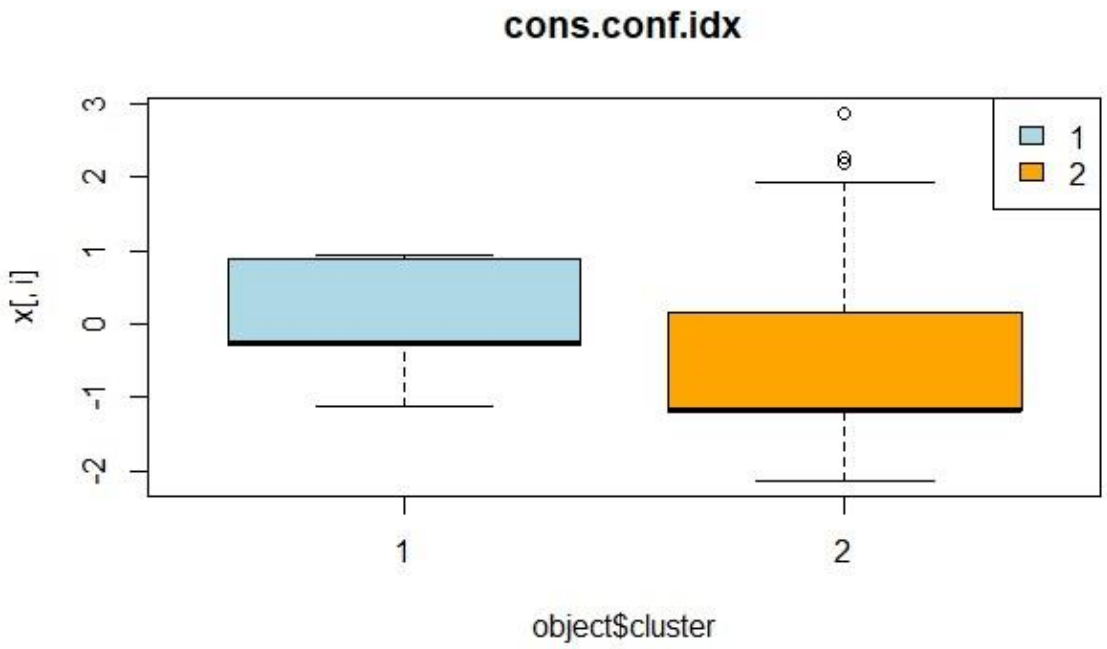

pdays
0.0
0.2
0.4
0.6
0.8
last week
last 2 weeks
last 3 weeks
No contacted before

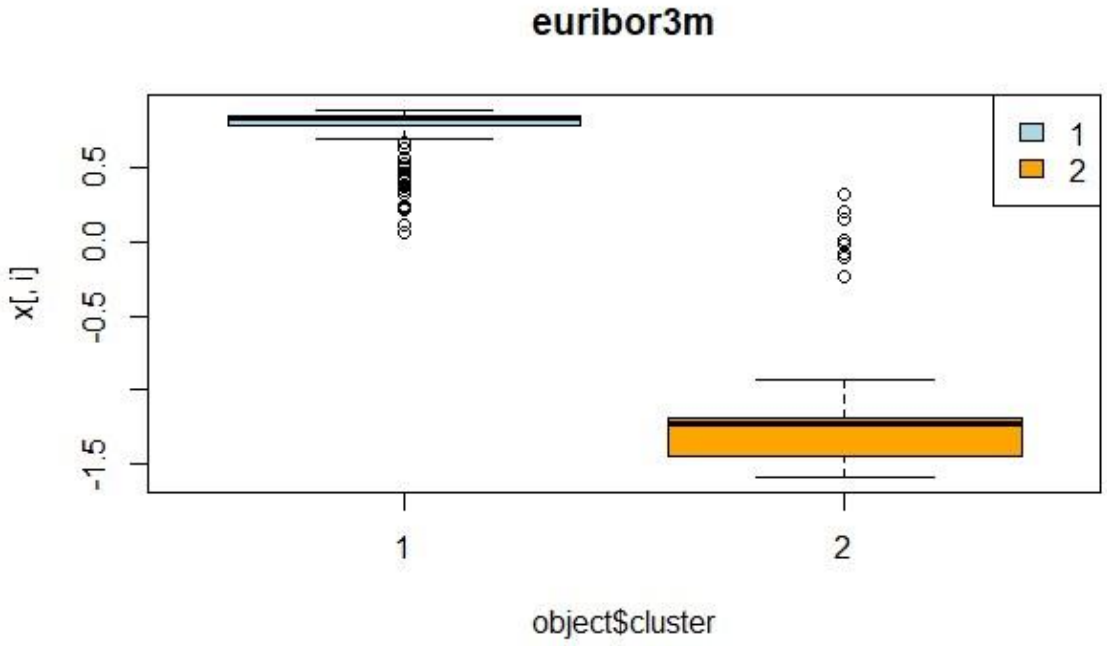

previous
x[, i]
0
2
4
6
8
10
1
2
object$cluster
1
2

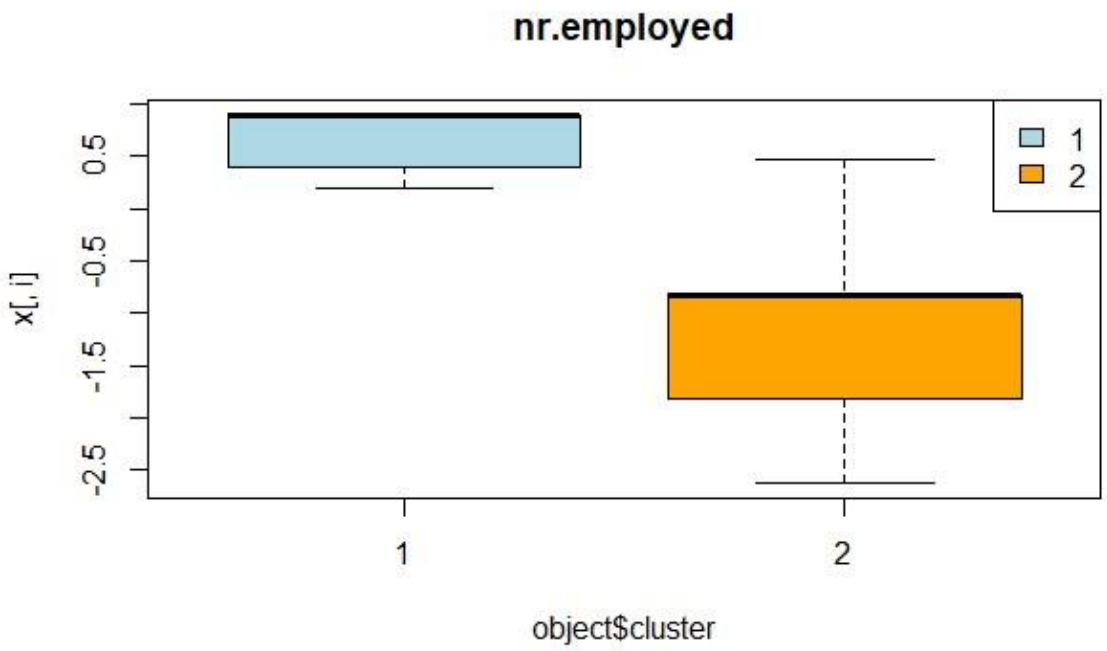

poutcome
0.0
0.2
0.4
0.6
0.8
failure
nonexistent
success

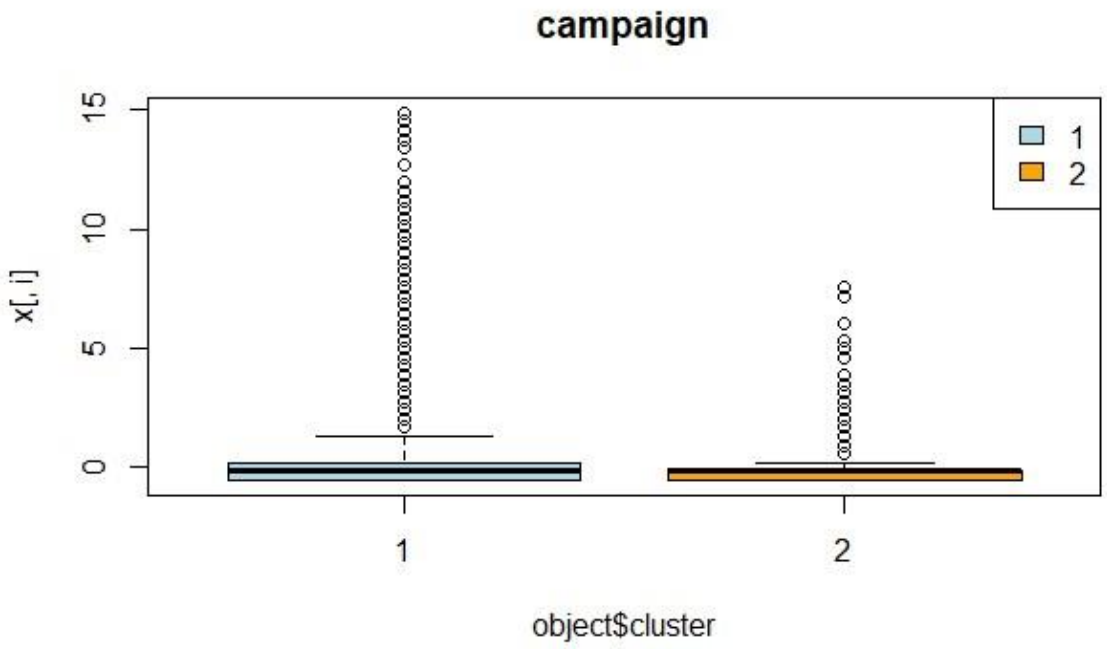

emp.var.rate
x[, i]
1.0
0.0
-1.0
-2.0
1
2
object$cluster
1
2


cons.price.idx
x[, i]
2
1
0
-1
-2
1
2
object$cluster
1
2


cons.conf.idx
x[, i]
3
2
1
0
-1
-2
1
2
object$cluster


euribor3m
x[, i]
0.5
0.0
-0.5
-1.5
1
2
1
2
object$cluster


nr.employed
x[, i]
0.5
-0.5
-1.5
-2.5
1
2
object$cluster
1
2


campaign
x[, i]
15
10
5
0
1
2
object$cluster
1
2