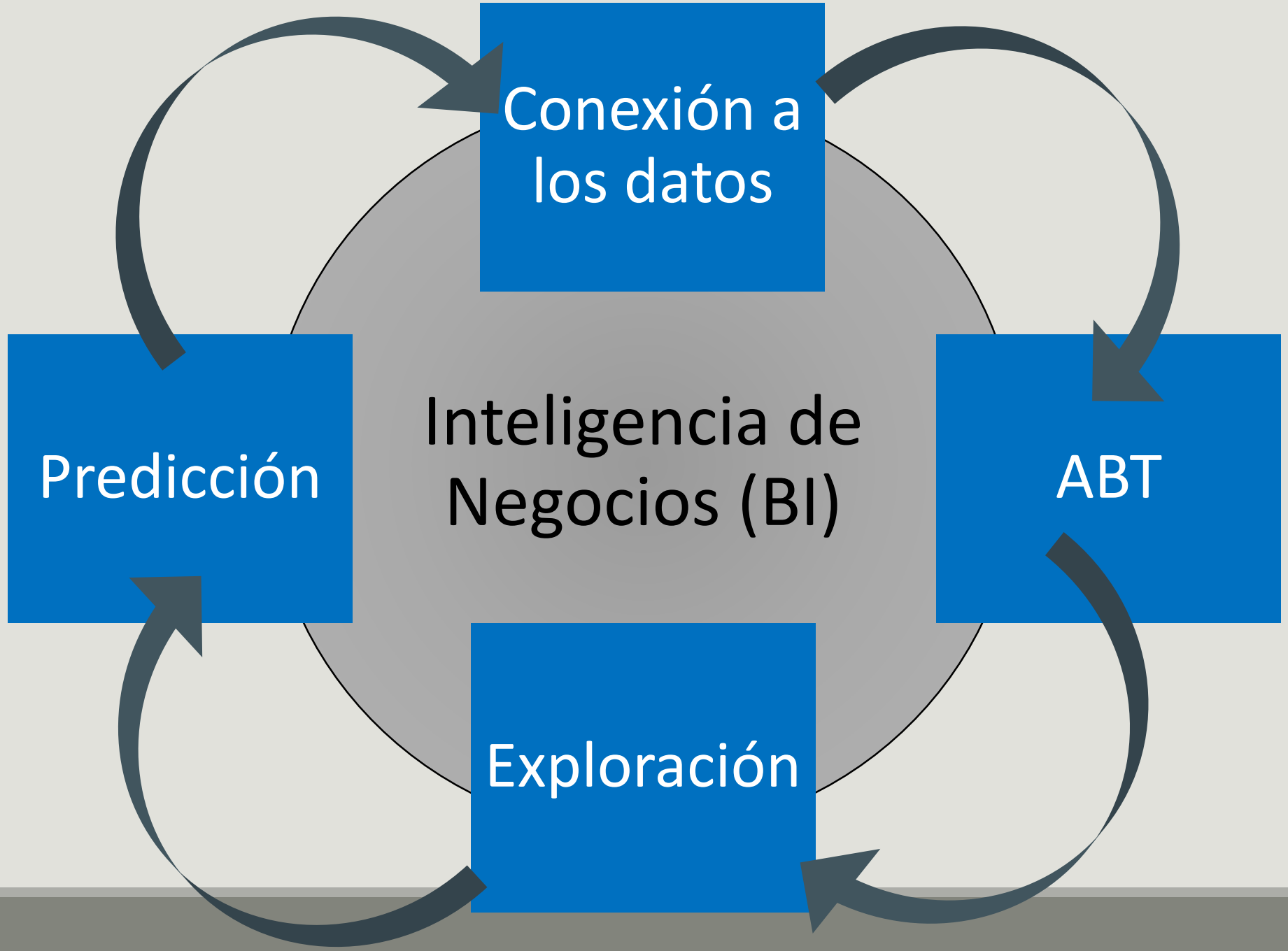
Aplicaciones de Machine Learning en la industria del retail

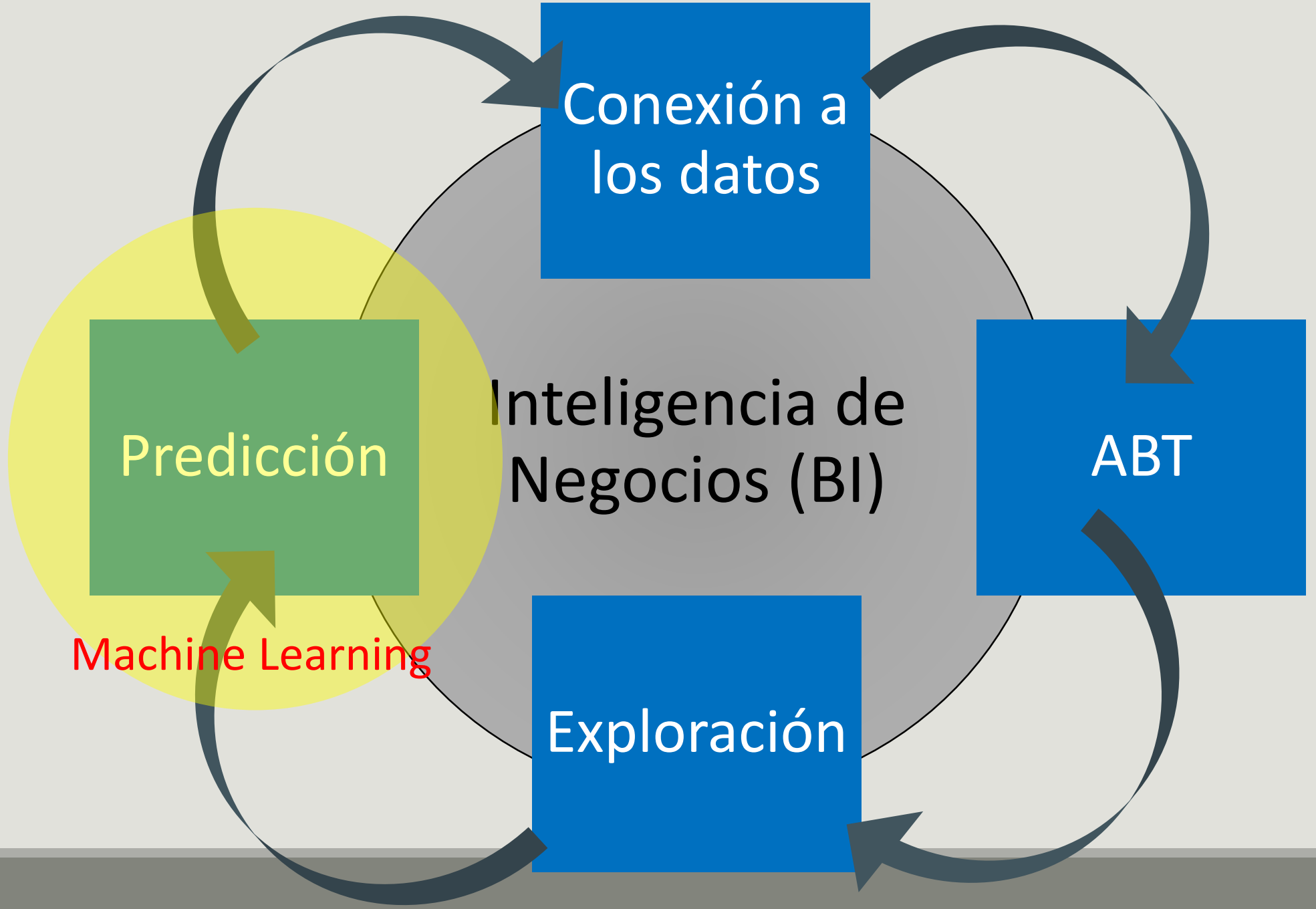
Rafael Ascanio

Senior Data Scientist

Rafael Ascanio

- Master of Science en Estadística con más de 18 años de experiencia en procesamiento de datos con miras a optimización de procesos de manufactura y servicios
- Académico de pre y postgrado en Estadística, Diseño Experimental, Análisis de Regresión, Análisis Multivariado y Modelos de Minería de Datos
- Líder Técnico para implementación de soluciones de riesgo en instituciones financieras
- Data Scientist


Conexión a los datos
ABT
Exploración
Predicción
Inteligencia de Negocios (BI)


Conexión a los datos
ABT
Exploración
Predicción
Inteligencia de Negocios (BI)
Machine Learning

¿Qué es el aprendizaje automático de máquinas o Machine Learning?

- ✓ Un término muy de moda
- ✓ No significa que las máquinas pasarán a hacer el trabajo de las personas
- ✓ Tampoco representa un robot haciendo labores complejas
- ✓ No es “Rocket Science”

Aprendizaje automático de máquinas (Machine Learning)

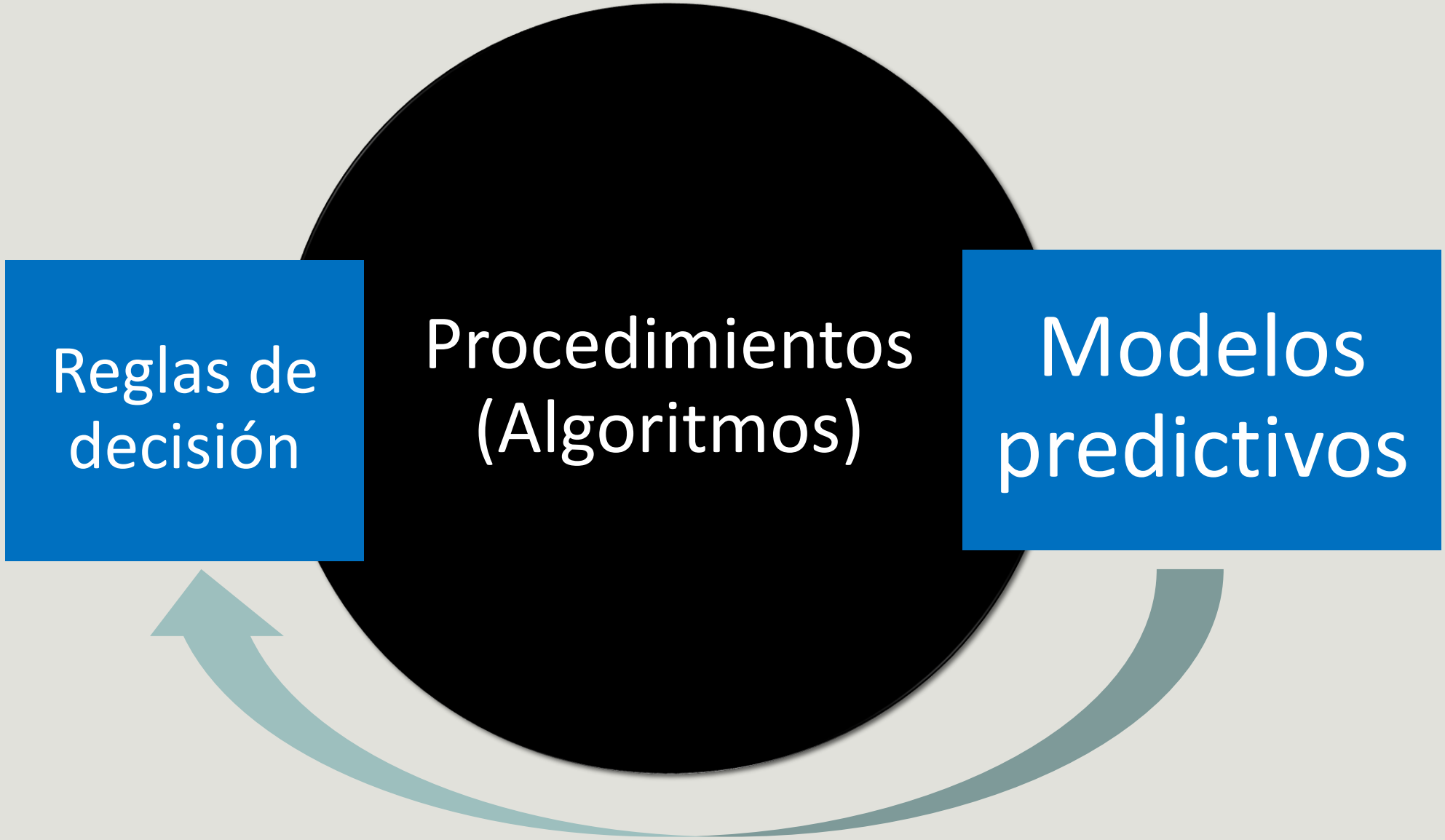
Se trata de crear reglas de decisión con base en la observación de ejemplos

Aprendizaje automático de máquinas (Machine Learning)

Procedimientos con más de 100 años de antigüedad

Capacidades de procesamiento de la información permiten la ejecución de tareas más complejas en menor tiempo

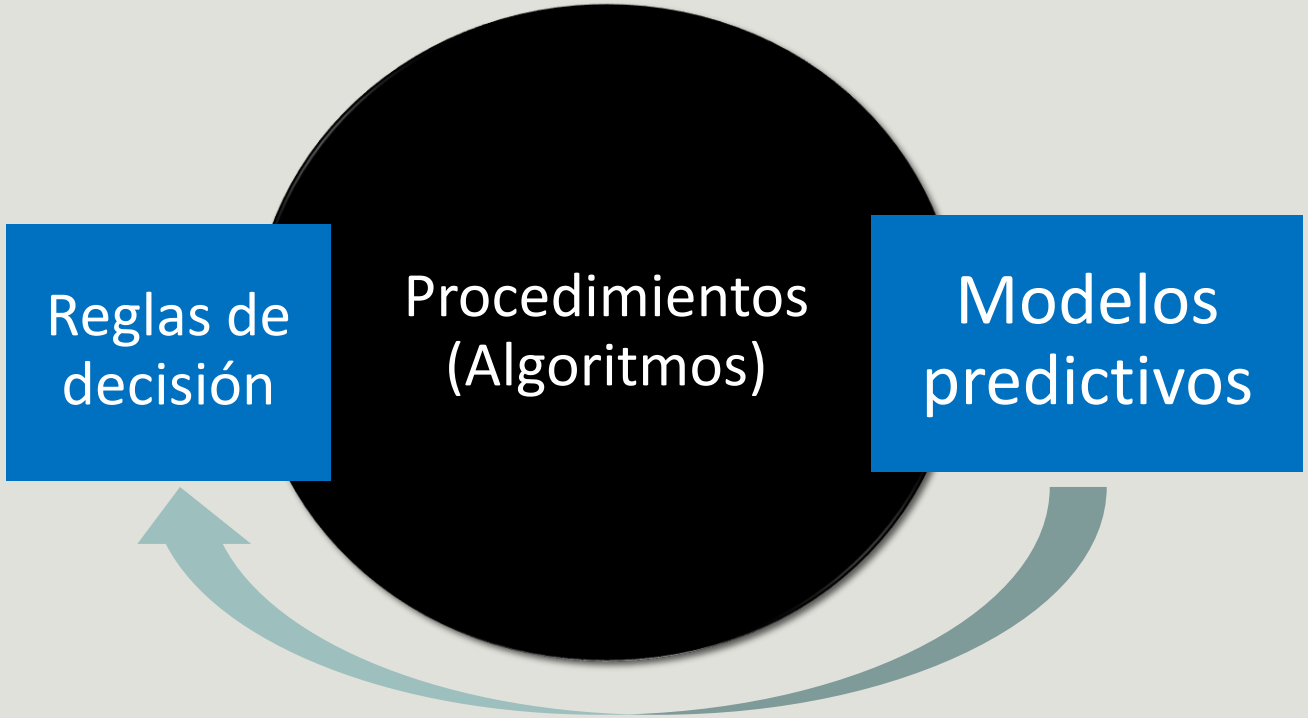
Grandes volúmenes de información permiten el desarrollo de reglas de decisión cada vez más exactas y precisas


Reglas de decisión
Procedimientos (Algoritmos)
Modelos predictivos
Medición del riesgo (Incertidumbre)

Confianza en la predicción

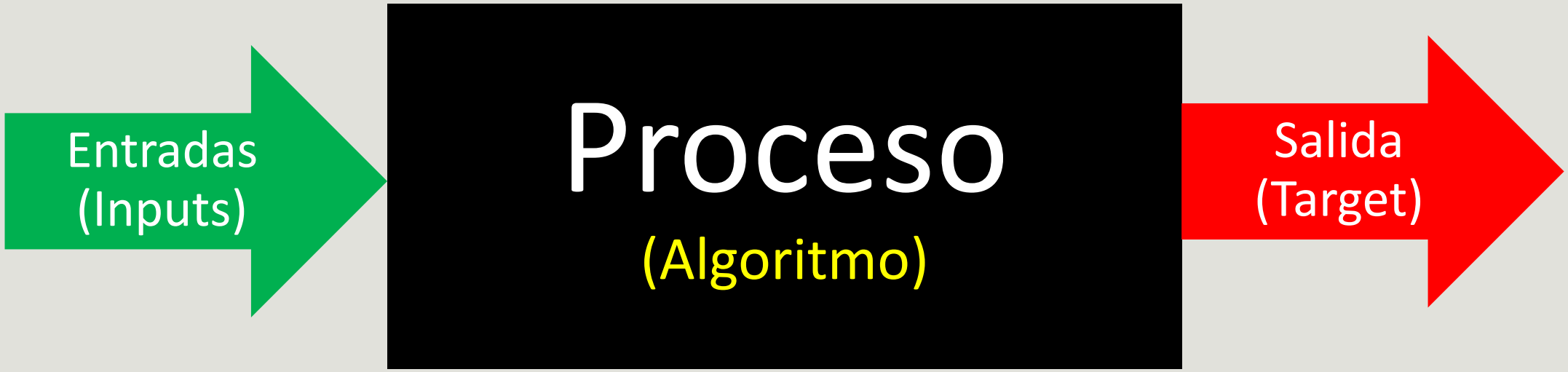
% Confianza = 100% - % Incertidumbre

Desercion	Nivel Confianza Predi...	Nombre	Apellido	Edad	PerfilViaje	DailyRate	Departamento	DistanciaHogar	NivelEducación	Especialidad
No	0.69	Benjamin	Piper	52	Travel_Rarely	258	Research & Development	8	4	Other
No	0.88	Felicity	Russell	37	Travel_Rarely	1462	Research & Development	11	3	Medical
No	0.94	Jake	Newman	35	Travel_Frequently	200	Research & Development	18	2	Life Sciences
No	0.67	Luke	Powell	25	Travel_Rarely	949	Research & Development	1	3	Technical Degree
No	0.64	Deirdre	Buckland	26	Travel_Rarely	652	Research & Development	7	3	Other
Yes	0.6	Victor	Stewart	29	Travel_Rarely	332	Human Resources	17	3	Other
No	0.7	Lisa	Piper	49	Travel_Frequently	1475	Research & Development	28	2	Life Sciences
No	0.57	Evan	Brown	29	Travel_Frequently	337	Research & Development	14	1	Other



"Los modelos predictivos, hoy día son herramientas de negocios esenciales que, de la mano con el buen juicio de los expertos, permitiría a las organizaciones alcanzar las estrellas"

Construcción de un modelo predictivo



Entradas del modelo (Inputs)

Datos estructurados

Modelos predictivos

Modelos lineales	Modelos de clasificación
Salida	Salida
Numérica	Categórica

Algoritmos de uso frecuente

Modelos lineales	Modelos de clasificación	Optimización
Regresiones lineales	Árboles de decisión	Algoritmos de optimización
Series temporales	Regresiones logísticas	
	Naive Bayes	
	Redes neuronales	
	Bosques aleatorios	
	Máquinas de Soporte Vectorial	
	Conglomerados	

¿Qué hacen estos algoritmos?

Modelos lineales

Regresiones lineales →

- Generan una expresión en forma de polinomio que relaciona las entradas con la salida
- Útiles para predecir salidas numéricas no afectas a estacionalidad y/o ciclo

---

Series temporales →

- Salida producida por su relación con las entradas, estando ordenadas caso a caso de manera cronológica
- Útiles para predecir demanda, precios de venta al menor y cualquier otra salida afecta a estacionalidad y/o ciclo

¿Qué hacen estos algoritmos?

Modelos de clasificación

Árboles de decisión

- Reglas de decisión por jerarquías
- Comienza con el input que más peso tiene en la predicción de la salida

- Visión completa de cómo interviene cada input (y en que orden deben ser considerados) en la selección de una categoría de la salida
- Útil para definir la ruta a seguir para evaluar si un cliente tiene la condición necesaria o no para la concesión de un crédito de consumo
- La ruta sería la secuencia de reglas de decisión por jerarquías

¿Qué hacen estos algoritmos?

Modelos de clasificación

- Regresiones logísticas
- Naive Bayes
- Redes neuronales
- Bosques aleatorios
- Máquinas de Soporte Vectorial

Generan salidas categóricas:

- Si/No
- Bueno/Malo
- Aprobado/Rechazado

Predicción de eventos que pueden tener dos o más categorías:

- Predicción del comportamiento de clientes (Behaviour)
- Predicción de fuga (Churn/Attrition)
- Selección de nuevos clientes (Acquisition)
- Selección de candidatos para una oferta
- Predicción de fallos en sistemas

¿Qué hacen estos algoritmos?

Modelos de clasificación

Conglomerados

Construyen grupos de casos similares en torno a todos los inputs de manera simultánea

- Segmentación de clientes
- Reconocimiento de familias de productos con estacionalidad similar
- Agrupamiento de clientes con similar condición de riesgo
- Reconocimiento del perfil de clientes potenciales

¿Qué hacen estos algoritmos?

Optimización

Algoritmos de optimización

Se produce un “set” de inputs en torno a una salida máxima o mínima

- Máximo revenue
- Mínimo costo

Usos de los algoritmos en retail

Caso de uso	Algoritmos	Fuentes
Gestión avanzada del inventario	Regresión lineal	Datos transaccionales
Gestión de la demanda	Modelos de clasificación	Repositorio de datos
Valor del cliente	Series temporales	Redes sociales
Probabilidad de respuesta de una campaña	Conglomerados	Sensores
Estrategia de optimización de precios	Optimización	Motores de decisión
Segmentación de clientes		Fotografías
Patrones de comportamiento de clientes		Texto
Marketing personalizado		
Analítica de Crédito y de las tarjetas de fidelidad		
Próxima mejor oferta		
Cross sell/Up sell		
Apertura de próxima sucursal		

El uso de estos algoritmos permitirá administrar el inventario con miras a minimizar el tiempo de permanencia de los artículos en stock, pero garantizando la existencia en todo momento:

- Modelamiento de la demanda en el tiempo
- Optimización de costos

Usos de los algoritmos en retail

Caso de uso	Algoritmos	Fuentes
Gestión avanzada del inventario	Regresión lineal	Datos transaccionales
Gestión de la demanda	Modelos de clasificación	Repositorio de datos
Valor del cliente	Series temporales	Redes sociales
Probabilidad de respuesta de una campaña	Conglomerados	Sensores
Estrategia de optimización de precios	Optimización	Motores de decisión
Segmentación de clientes		Fotografías
Patrones de comportamiento de clientes		Texto
Marketing personalizado		
Analítica de Crédito y de las tarjetas de fidelidad		
Próxima mejor oferta		
Cross sell/Up sell		
Apertura de próxima sucursal		

Usos de los algoritmos en retail

Caso de uso	Algoritmos	Fuentes
Gestión avanzada del inventario	Regresión lineal	Datos transaccionales
Gestión de la demanda	Modelos de clasificación	Repositorio de datos
Valor del cliente	Series temporales	Redes sociales
Probabilidad de respuesta de una campaña	Conglomerados	Sensores
Estrategia de optimización de precios	Optimización	Motores de decisión
Segmentación de clientes		Fotografías
Patrones de comportamiento de clientes		Texto
Marketing personalizado		
Analítica de Crédito y de las tarjetas de fidelidad		
Próxima mejor oferta		
Cross sell/Up sell		
Apertura de próxima sucursal		

Usos de los algoritmos en retail

Caso de uso	Algoritmos	Fuentes
Gestión avanzada del inventario	Regresión lineal	Datos transaccionales
Gestión de la demanda	Modelos de clasificación	Repositorio de datos
Valor del cliente	Series temporales	Redes sociales
Probabilidad de respuesta de una campaña	Conglomerados	Sensores
Estrategia de optimización de precios	Optimización	Motores de decisión
Segmentación de clientes		Fotografías
Patrones de comportamiento de clientes		Texto
Marketing personalizado		
Analítica de Crédito y de las tarjetas de fidelidad		
Próxima mejor oferta		
Cross sell/Up sell		
Apertura de próxima sucursal		

Usos de los algoritmos en retail

Caso de uso	Algoritmos	Fuentes
Gestión avanzada del inventario	Regresión lineal	Datos transaccionales
Gestión de la demanda	Modelos de clasificación	Repositorio de datos
Valor del cliente	Series temporales	Redes sociales
Probabilidad de respuesta de una campaña	Conglomerados	Sensores
Estrategia de optimización de precios	Optimización	Motores de decisión
Segmentación de clientes		Fotografías
Patrones de comportamiento de clientes		Texto
Marketing personalizado		
Analítica de Crédito y de las tarjetas de fidelidad		
Próxima mejor oferta		
Cross sell/Up sell		
Apertura de próxima sucursal		

Uso de un algoritmo de Machine Learning para entrenar un modelo de selección binaria

Caso: Modelo de propensión a compra

La idea

Objetivo

Proponer una manera simple de focalizar el tiempo y los esfuerzos para captar la atención sólo de aquellos que tienen interés en las ofertas del negocio

Método

Crear una regla de decisión para enviar campañas de productos y servicios a quienes tienen mayor probabilidad de responder de manera positiva a la invitación a compra

Insumo

Datos contenidos en una tabla de información de clientes

El procedimiento

Objetivo

Utilizar un procedimiento de regresión logística para generar un modelo de selección binaria, capaz de clasificar a los clientes en dos categorías: quienes probablemente compren y quienes probablemente no compren.

Flujo del proceso

1. Carga de datos
2. Inspección de la distribución de los datos por categoría de la respuesta
3. Generación y balanceo de las submuestras para crear los conjuntos de entrenamiento y validación
4. Categorización de las variables de intervalo
5. Construcción de la ecuación de regresión logística
6. Evaluación del desempeño del modelo
7. Despliegue del modelo sobre un nuevo conjunto de datos sin respuesta observada (Scoring)

Herramientas analíticas

- ✓ R versión 3.4.4 (2018-03-15)
- ✓ Oracle Data Visualization for Desktop

Flujo del proceso

1. Carga de datos:

inputData <- read.csv("D:/Descargas/Compra.csv")
head(inputData)

```
> inputData <- read.csv("D:/Descargas/Compra.csv")
> head(inputData)
  Edad        Sector     Id                    Educ Resid        Eciv
1   39       Publico  77516  Universitaria_Completa    13     Soltero  Admin
2   50  Autoempleado  83311  Universitaria_Completa    13      Casado
3   38       Privado 215646              Secundaria     9  Divorciado   Mant
4   53       Privado 234721                Primaria     7      Casado   Mant
5   28       Privado 338409  Universitaria_Completa    13      Casado    Esp
6   37       Privado 284582               Postgrado    14      Casado
```

2. Inspección de la distribución de los datos por categoría de la respuesta:

table(inputData$Compra)

```
> table(inputData$Compra)

    0     1
22654  7508
```

Flujo del proceso

3. Generación y balanceo de las submuestras para crear los conjuntos de entrenamiento y validación:

```
input_ones <- inputData[which(inputData$Compra == 1), ]
input_zeros <- inputData[which(inputData$Compra == 0), ]
set.seed(100)
input_ones_training_rows <- sample(1:nrow(input_ones), 0.7*nrow(input_ones))
input_zeros_training_rows <- sample(1:nrow(input_zeros), 0.7*nrow(input_ones))
training_ones <- input_ones[input_ones_training_rows, ]
training_zeros <- input_zeros[input_zeros_training_rows, ]
trainingData <- rbind(training_ones, training_zeros)
test_ones <- input_ones[-input_ones_training_rows, ]
test_zeros <- input_zeros[-input_zeros_training_rows, ]
testData <- rbind(test_ones, test_zeros)
```

Flujo del proceso

View(trainingData)
View(testData)

Data: trainingData

	row.names	Edad	Sector	Id	Educ	Resid	Eciv	Ocup
1	9429	55	Autoempleado	124137	Universitaria_Incompleta	10	Casado	Ventas
2	7842	35	Privado	66297	Universitaria_Completa	13	Soltero	Especialista
3	16819	37	Privado	278576	Secundaria	9	Casado	Operador_Maquina
4	1641	35	Publico	103260	Postgrado	14	Casado	Especialista
5	14261	59	Privado	98361	Universitaria_Incompleta	10	Casado	Mecanica
6	14651	52	Privado	284329	Universitaria_Incompleta	10	Casado	Mecanica
7	24597	32	Autoempleado	124919	Universitaria_Completa	13	Casado	Otros
8	11277	52	Privado	168553	Secundaria	9	Casado	Mecanica
9	16631	58	Privado	156040	Tecnica	12	Casado	Operador_Maquina
10	5099	27	Privado	292472	Universitaria_Incompleta	10	Casado	Especialista
11	19033	57	Publico	140711	Universitaria_Completa	13	Casado	Especialista
12	26697	30	Autoempleado	116666	Postgrado	14	Divorciado	Especialista
13	8469	38	Privado	127601	Universitaria_Incompleta	10	Viudo	Ejecutivo
14	12131	60	Privado	198170	Universitaria_Completa	13	Casado	Ejecutivo
15	23085	28	Privado	130067	Universitaria_Completa	13	Casado	Especialista
16	20383	56	Privado	188856	Universitaria_Completa	13	Soltero	Ejecutivo
17	6122	52	Publico	338816	Postgrado	14	Soltero	Administracion
18	10900	41	Privado	204410	Secundaria	9	Casado	Mecanica
19	10944	49	Privado	122385	Postgrado	14	Casado	Ejecutivo
20	20979	53	Autoempleado	284329	Postgrado	14	Divorciado	Ejecutivo
21	16279	38	Privado	312271	Universitaria_Completa	13	Casado	Especialista
22	21633	70	Autoempleado	37203	Postgrado	14	Casado	Especialista
23	16354	28	Privado	51461	Universitaria_Completa	13	Casado	Administracion
24	22638	35	Privado	225860	Tecnica	11	Casado	Mecanica
25	12765	53	Privado	47396	Universitaria_Incompleta	10	Casado	Administracion
26	5118	44	Privado	169397	Secundaria	9	Casado	Mecanica

Flujo del proceso

4. Categorización de los atributos y las variables de intervalo.

   Cálculo del Information Value (IV)

```
library(smbinning)
factor_vars <- c ("Sector", "Educ", "Ocup", "Eciv", "Ubic", "Region", "Genero", "Ciudad")
continuous_vars <- c("Edad", "Resid", "Fijo", "Variable", "Distancia")
iv_df <- data.frame(VARS=c(factor_vars, continuous_vars), IV=numeric(13))

for(factor_var in factor_vars){
  smb <- smbinning.factor(trainingData, y="Compra", x=factor_var)
  if(class(smb) != "character"){
    iv_df[iv_df$VARS == factor_var, "IV"] <- smb$iv
  }
}
```

Flujo del proceso

```
for(continuous_var in continuous_vars){
  smb <- smbinning(trainingData, y="Compra",
x=continuous_var)
  if(class(smb) != "character"){
    iv_df[iv_df$VARS == continuous_var, "IV"] <- smb$iv
  }
}

iv_df <- iv_df[order(-iv_df$IV), ]
iv_df
```

```
> iv_df <- iv_df[order(-iv_df$IV), ]
> iv_df
          VARS      IV
5         Ubic 1.4640
4         Eciv 1.2926
9         Edad 1.0872
11        Fijo 0.7796
10      Resid 0.7485
2         Educ 0.7337
13 Distancia 0.4278
7       Genero 0.3093
12   Variable 0.1598
1       Sector 0.0784
6       Region 0.0745
8       Ciudad 0.0271
3         Ocup 0.0000
```

Flujo del proceso

5. Construcción de la ecuación de regresión logística:

```
logitMod <- glm(Compra ~ Ubic + Eciv + Edad + Fijo + Resid + Educ + Distancia + Genero + Variable +
Sector + Region + Ciudad + Ocup, data=trainingData, family=binomial(link="logit"))
predicted <- plogis(predict(logitMod, testData))  # predicted scores
# or
predicted <- predict(logitMod, testData, type="response")

library(InformationValue)
optCutOff <- optimalCutoff(testData$Compra, predicted)[1]
```

summary(logitMod)

```
> summary(logitMod)

Call:
glm(formula = Compra ~ Ubic + Eciv + Edad + Fijo + Resid + Educ +
    Distancia + Genero + Variable + Sector + Region + Ciudad +
    Ocup, family = binomial(link = "logit"), data = trainingData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.2701  -0.5146   0.0000   0.6139   3.4115

Coefficients:
                              Estimate    Std. Error z value             Pr(>|z|)
(Intercept)                -4.88696953    1.45876592  -3.350             0.000808 ***
Ubic Centro                 1.07310398    0.19752274   5.433 0.00000005547273947 ***
Ubic Estacion               0.97382081    0.23075587   4.220 0.00002441565796133 ***
Ubic Mall Tipo A            0.32829302    0.37098641   0.885             0.376200
Ubic Mall Tipo B            1.64127214    0.38663980   4.245 0.00002186282734809 ***
Ubic Otro                   0.34584774    0.34712445   0.996             0.319094
```

Flujo del proceso

6. Evaluación del desempeño del modelo:

misClassError(testData$Compra, predicted, threshold = optCutOff)
plotROC(testData$Compra, predicted)
Concordance(testData$Compra, predicted)
sensitivity(testData$Compra, predicted, threshold = optCutOff)
specificity(testData$Compra, predicted, threshold = optCutOff)
confusionMatrix(testData$Compra, predicted, threshold = optCutOff)

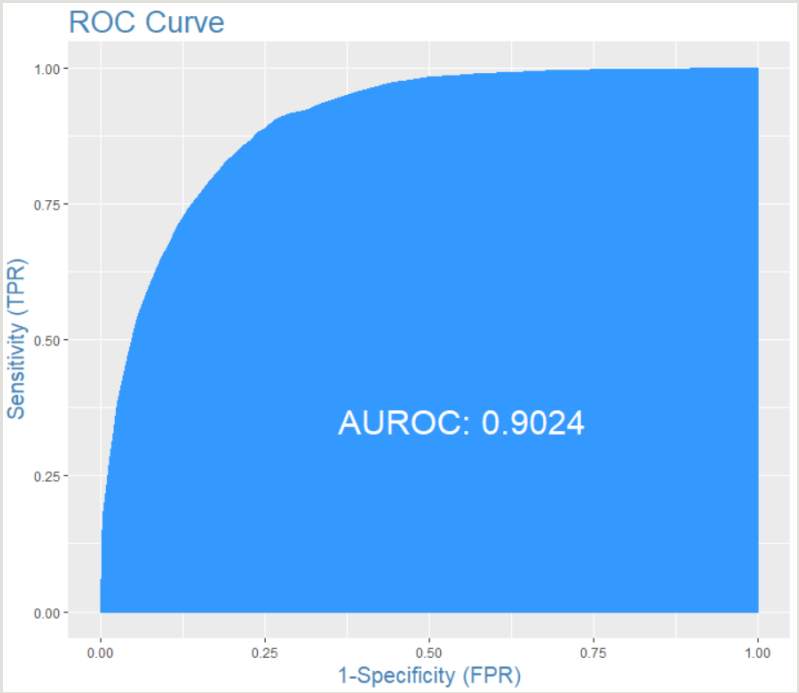
```
> misClassError(testData$Compra, predicted, threshold = optCutOff)
[1] 0.0933
>
> plotROC(testData$Compra, predicted)
>
> Concordance(testData$Compra, predicted)
$Concordance
[1] 0.9025358

$Discordance
[1] 0.09746421

$Tied
[1] 0.000000000000000001387779

$Pairs
[1] 39199947

>
> sensitivity(testData$Compra, predicted, threshold = optCutOff)
[1] 0.3067022
> specificity(testData$Compra, predicted, threshold = optCutOff)
[1] 0.9844244
>
> confusionMatrix(testData$Compra, predicted, threshold = optCutOff)
      0    1
0 17128 1562
1   271  691
```



Flujo del proceso

7. Almacenamiento del modelo:

   save(file="modelo_Compra", logitMod)

8. Despliegue del modelo sobre un nuevo conjunto de datos sin respuesta observada (Scoring):

   scoreData <- read.csv("D:/Descargas/scoreData.csv")
   head(scoreData)

   load(file="modelo_Compra")

   scoreData$pred <- predict(logitMod, newdata=scoreData, type='response')

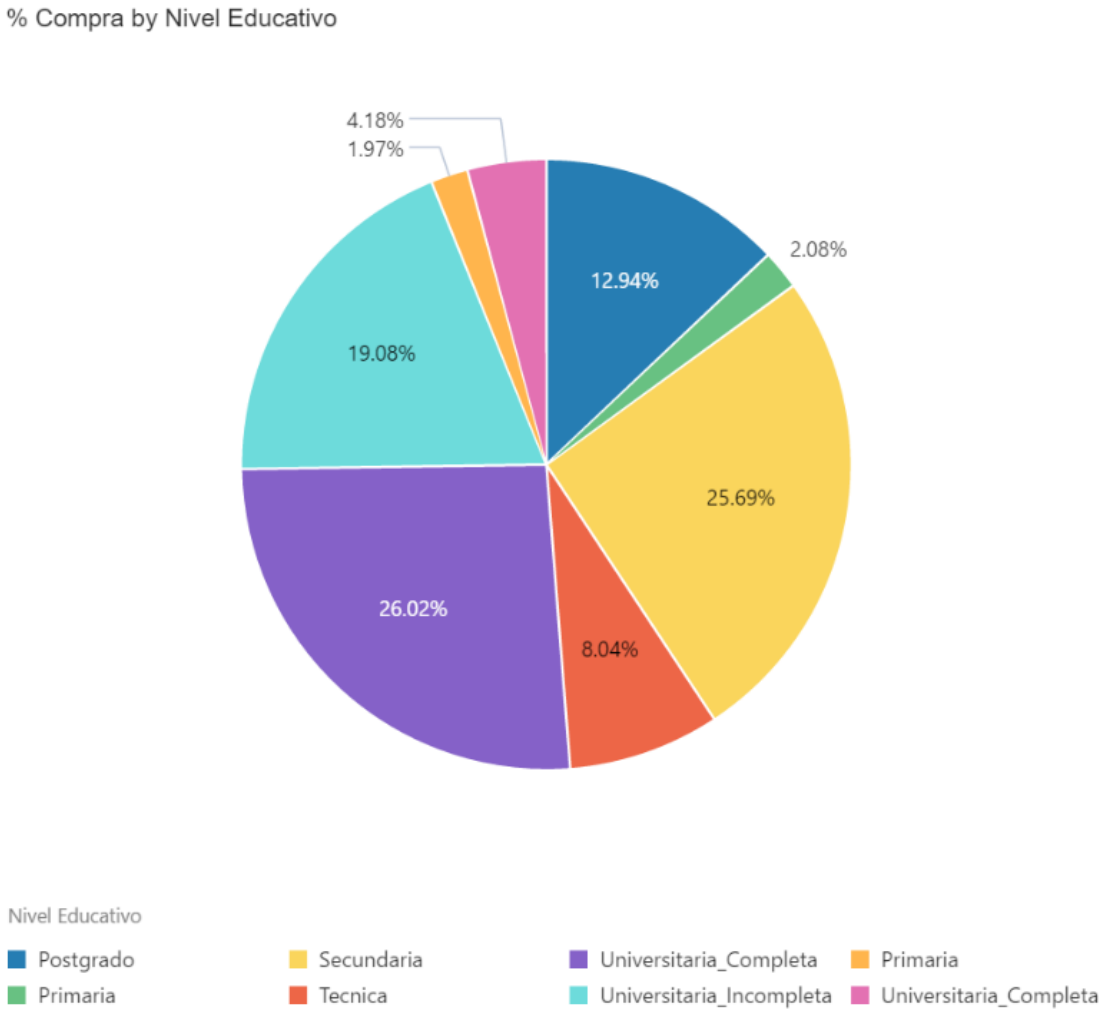
```
> scoreData <- read.csv("D:/Descargas/scoreData.csv")
> head(scoreData)
  Edad        Sector     Id                    Educ Resid        Eciv
1   48  Autoempleado 191277                Postgrado    16      Casado  Especi
2   37       Privado 202683  Universitaria_Incompleta    10      Casado
3   48       Privado 171095                  Tecnica    12  Divorciado     Eje
4   32       Publico 249409               Secundaria     9     Soltero
5   76       Privado 124191                Postgrado    14      Casado     Eje
6   44       Privado 198282   Universitaria_Completa    13      Casado     Eje
```

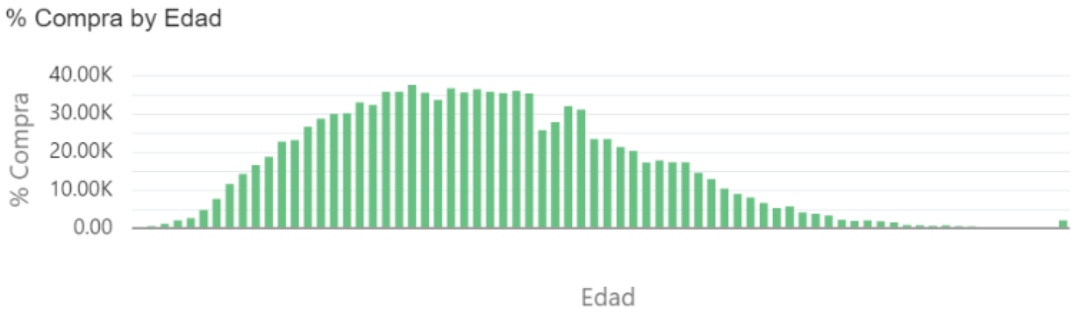
Flujo del proceso

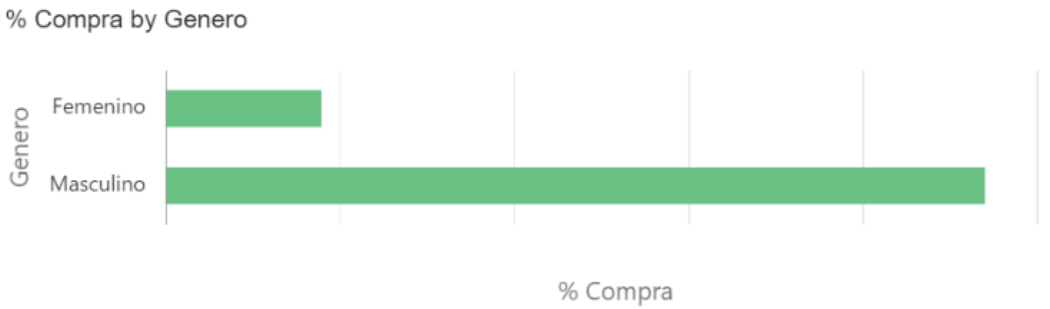
Salida del modelo (Output)

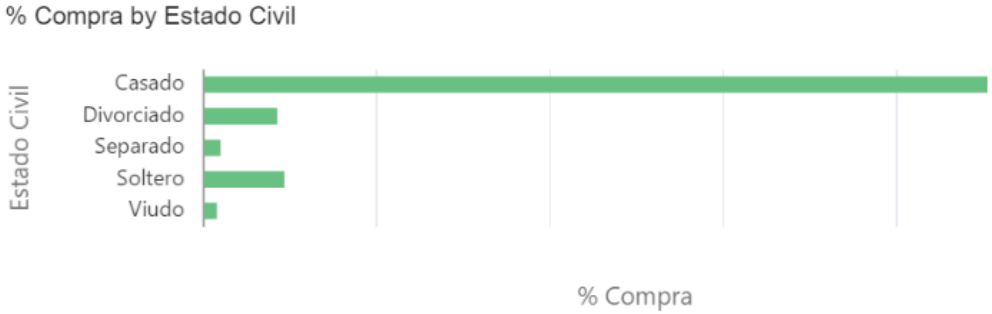
	Edad	Sector	Id	Educ	Resid	Eciv	Ocup	Ubic	Region	Genero	Fijo	Variable	Distancia	Ciudad	pred
1	48	Autoempl	191277	Postgrado	16	Casado	Especialist	Mall Tipo	Oriente	Masculinc	0	1902	60	Capital	0.986603
2	37	Privado	202683	Universita	10	Casado	Ventas	Mall Tipo	Oriente	Masculinc	0	0	48	Capital	0.658682
3	48	Privado	171095	Tecnica	12	Divorciad	Ejecutivo	Estacion	Oriente	Femenino	0	0	40	Region_VI	0.177025
4	32	Publico	249409	Secundari	9	Soltero	Otros	Avenida/C	Oeste	Masculinc	0	0	40	Capital	0.010485
5	76	Privado	124191	Postgrado	14	Casado	Ejecutivo	Mall Tipo	Oriente	Masculinc	0	0	40	Capital	0.955598
6	44	Privado	198282	Universita	13	Casado	Ejecutivo	Mall Tipo	Oriente	Masculinc	15024	0	60	Capital	0.999476
7	47	Autoempl	149116	Postgrado	14	Soltero	Especialist	Centro	Oriente	Femenino	0	0	50	Capital	0.281714
8	20	Privado	188300	Universita	10	Soltero	Tecnologi	Avenida/C	Oriente	Femenino	0	0	40	Capital	0.017991
9	29	Privado	103432	Secundari	9	Soltero	Mecanica	Centro	Oriente	Masculinc	0	0	40	Capital	0.072494
10	32	Autoempl	317660	Secundari	9	Casado	Mecanica	Mall Tipo	Oriente	Masculinc	7688	0	40	Capital	0.880494
11	30	Privado	194901	Primaria	7	Soltero	Mantenim	Avenida/C	Oriente	Masculinc	0	0	40	Capital	0.00641
12	31	Publico	189265	Secundari	9	Soltero	Administra	Centro	Oriente	Femenino	0	0	40	Capital	0.03096
13	42	Privado	124692	Secundari	9	Casado	Mantenim	Mall Tipo	Oriente	Masculinc	0	0	40	Capital	0.371125
14	24	Privado	432376	Universita	13	Soltero	Ventas	Otro	Oriente	Masculinc	0	0	40	Capital	0.112743
15	38	Privado	65324	Universita	15	Casado	Especialist	Mall Tipo	Oriente	Masculinc	0	0	40	Capital	0.921478
16	56	Autoempl	335605	Secundari	9	Casado	Otros	Mall Tipo	Oriente	Masculinc	0	1887	50	Region_IX	0.721988
17	28	Privado	377869	Universita	10	Casado	Ventas	Mall Tipo	Oriente	Femenino	4064	0	25	Capital	0.777095
18	36	Privado	102864	Secundari	9	Soltero	Operador_	Avenida/C	Oriente	Femenino	0	0	40	Capital	0.008415
19	53	Privado	95647	Primaria	5	Casado	Mantenim	Mall Tipo	Oriente	Masculinc	0	0	50	Capital	0.234412
20	56	Autoempl	303090	Universita	10	Casado	Ventas	Mall Tipo	Oriente	Masculinc	0	0	50	Capital	0.750022
21	49	Publico	197371	Tecnica	11	Casado	Mecanica	Mall Tipo	Oeste	Masculinc	0	0	40	Capital	0.601616
22	55	Privado	247552	Universita	10	Casado	Ventas	Mall Tipo	Oriente	Masculinc	0	0	56	Capital	0.815873
23	22	Privado	102632	Secundari	9	Soltero	Mecanica	Centro	Oriente	Masculinc	0	0	41	Capital	0.061674

Nuevas vistas del negocio (a partir de la salida del modelo)









“Essentially, all models are wrong...
but some are useful”

George Box

1919 – 2013

Rafael Ascanio
Email: rafael.ascanio@oracle.com
Celular: +56 9 3242 1772
Rafael Ascanio
https://www.linkedin.com/in/ascanioe