Regresión Logística vs Árboles de Decisión

Ejemplo con prueba de selección para Jefe de Analítica de Clientes

Análisis Multivariado Enfocado a la Gestión de Riesgos

Edith Johana Medina Hernández – edithjoh@gmail.com

Ejemplo con prueba de selección para Jefe de Analítica de Clientes

Como jefe de analítica de cliente usted debe comprender el comportamiento de los clientes, segmentarlos, estar en capacidad de presentarle a la alta gerencia de la compañía sus hallazgos y emplear modelos analíticos para determinar qué campañas se deberán hacer para mejorar la rentabilidad de los clientes.

Para el siguiente caso, usted deberá trabajar con datos reales de unas campañas se telemercadeo de un banco Portugués. El objetivo del ejercicio es analizar la base de clientes, estimar un modelo de predicción de **clientes que deben ser seleccionados en cada campaña** y con base en este proponer la estrategia sobre la cual se deben realizar las campañas de este producto.

La base de datos la puede descargar de http://archive.ics.uci.edu/ml/machine-learning-databases/00222/bank.zip usar el conjunto de datos “bank-full.csv”. Este conjunto de datos está descrito en el artículo académico **Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology** (lo puede descargar de la dirección http://hdl.handle.net/1822/14838), se recomiendo leerlo para tener mayor entendimiento del conjunto de datos y el contexto bajo el cual fue construido.

Puede encontrar la descripción del archivo de datos en la página web del repositorio:
http://archive.ics.uci.edu/ml/machine-learning-databases/00222/bank.zip

Lo que se solicitó en la Prueba de Selección

El entregable de este ejercicio es una presentación de PowerPoint que cubra los siguientes aspectos:

Capítulo técnico (Asuma que lo presentaría ante un público técnico con amplios conocimientos estadísticos):

1. Análisis descriptivo de la población
2. Modelo de predicción seleccionado.
    a. Justificación de la metodología de modelamiento seleccionada
    b. Presentar métricas de clasificación (Área bajo curva ROC)

Capítulo de Negocio (Asuma que lo presentará ante gerentes de negocio, con un buen entendimiento de los clientes pero bajo conocimiento técnico)

1. Principales hallazgos y conclusiones
2. Propuesta de estrategia para seleccionar los clientes buscando el mayor impacto con el menor costo posible.

Aparte del documento (archivo de PowerPoint), la presentación se deberá hacer personalmente (o por video conferencia) con una duración máxima de 30 minutos. Ambos capítulos son igualmente importantes y tendrán un peso de 50% cada uno.

Proceso de Selección: Jefe Analítica de Clientes

Análisis de Caso: Campañas de Depósito a Plazo

Edith Johana Medina Hernández

Marzo de 2017

Contenido

1. **Análisis Técnico:**
    - Análisis Descriptivo
    - Modelación y Evaluación de Modelos
2. **Análisis Estratégico:**
    - 3D del Territorio Estratégico de Análisis
    - Conclusiones y Recomendaciones

Referencias y Anexos

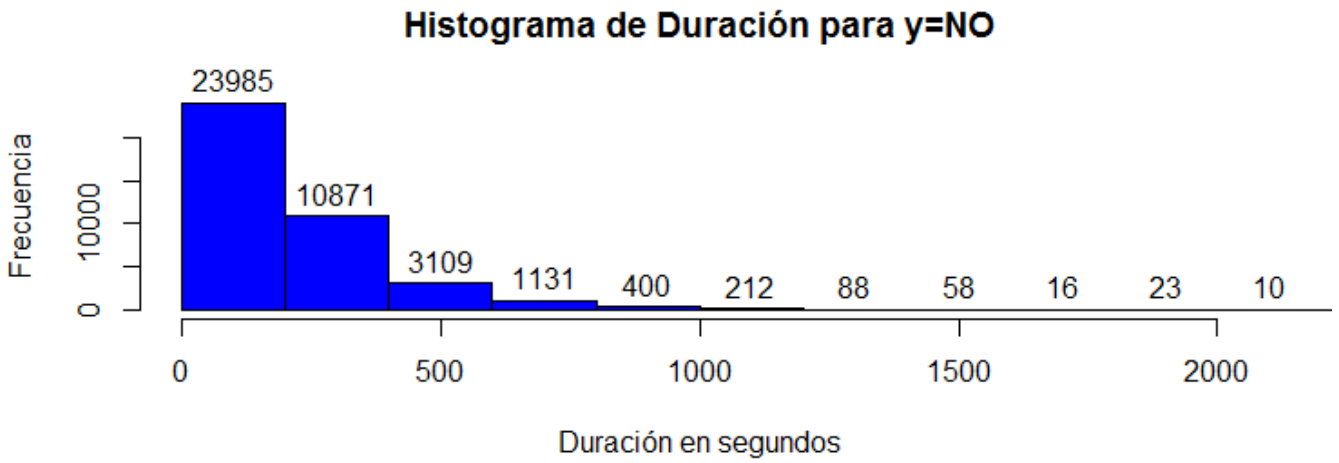
Análisis Técnico

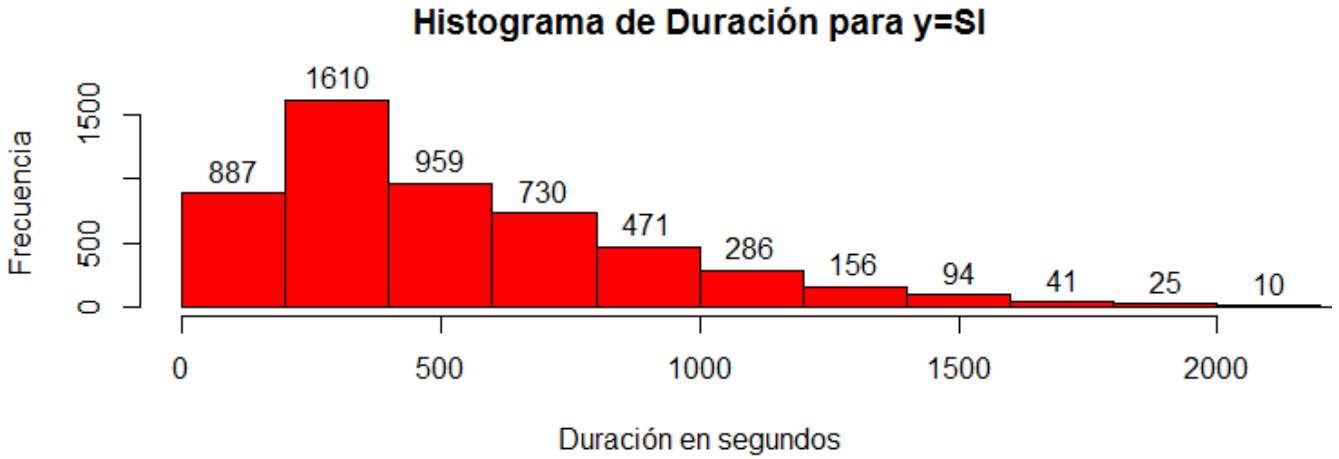
1. Análisis Descriptivo
2. Modelación

Variable: Duración de la llamada en el último contacto

Análisis Descriptivo

- La duración de la llamada en el contacto anterior para quienes se suscriben, suele ser mayor a 4 min
- Se observa que solo un 15% de las llamadas que en el anterior contacto duraron 3 minutos o más, no se suscriben





Estadísticos Descriptivos:

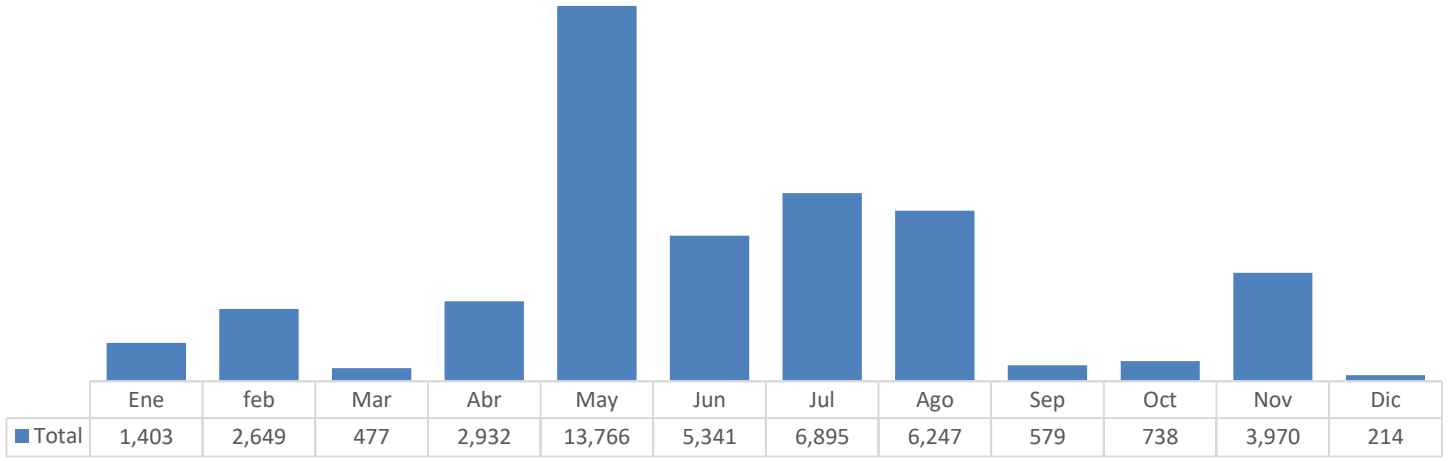
	NO se suscribe	
Indicador	Duración en segundos	Duración en Minutos
percentil 25	95	2
mediana	164	3
media	65	1 min 5 seg
percentil 75	70	1 min 10 seg
percentil 85	168	3

	Si se suscribe	
Indicador	Duración en segundos	Duración en Minutos
percentil 25	244	4
mediana	426	7
media	537	9
percentil 75	725	12

Variable: Mes de Contacto

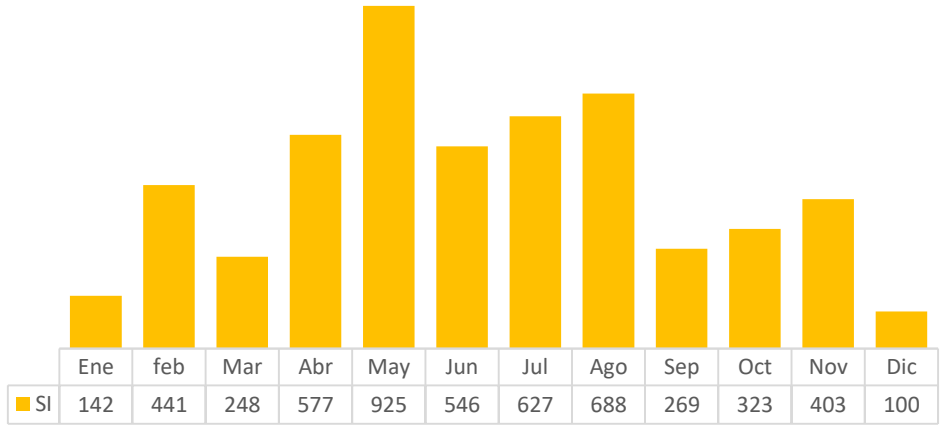
Análisis Descriptivo

Total de Contactos por Mes



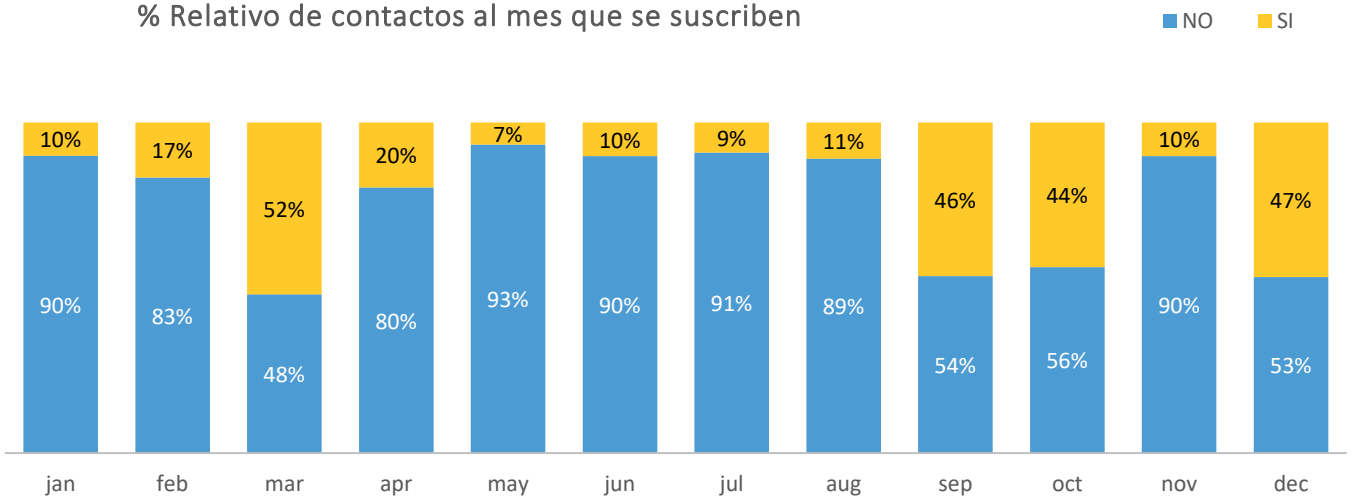
	Ene	feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
Total	1,403	2,649	477	2,932	13,766	5,341	6,895	6,247	579	738	3,970	214

Toral de suscripciones por mes



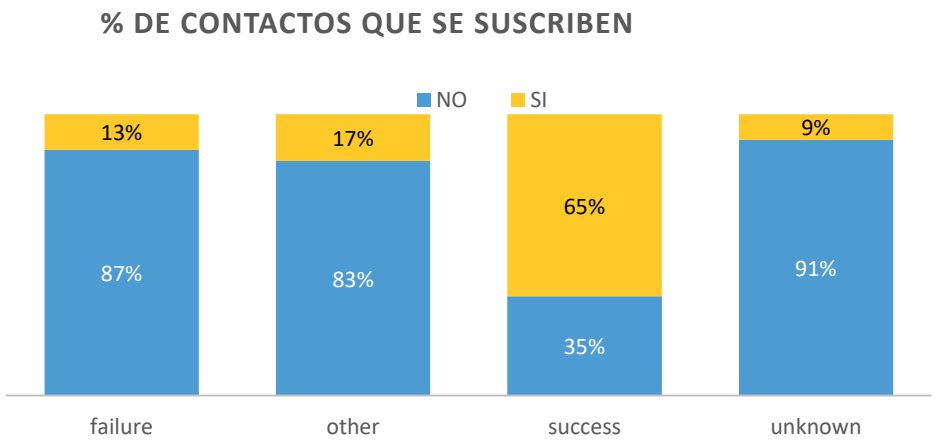
	Ene	feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
SI	142	441	248	577	925	546	627	688	269	323	403	100

% Relativo de contactos al mes que se suscriben

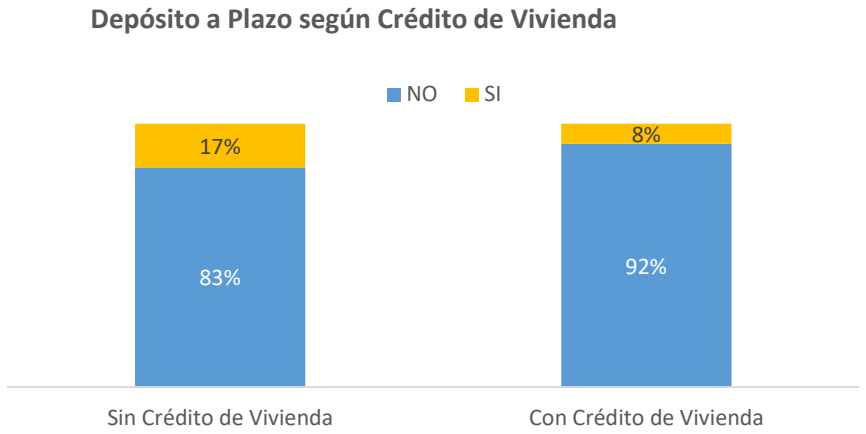


Los contactos realizados en los meses de Febrero, Marzo, Abril, Septiembre, Octubre y Diciembre, pese a ser inferiores en cantidad frente a los de otros meses, tienen mayor proporción de respuesta en SI

Variable Poutcome
Resultado de la campaña de marketing anterior



Variable Housing
Tiene préstamo de vivienda?



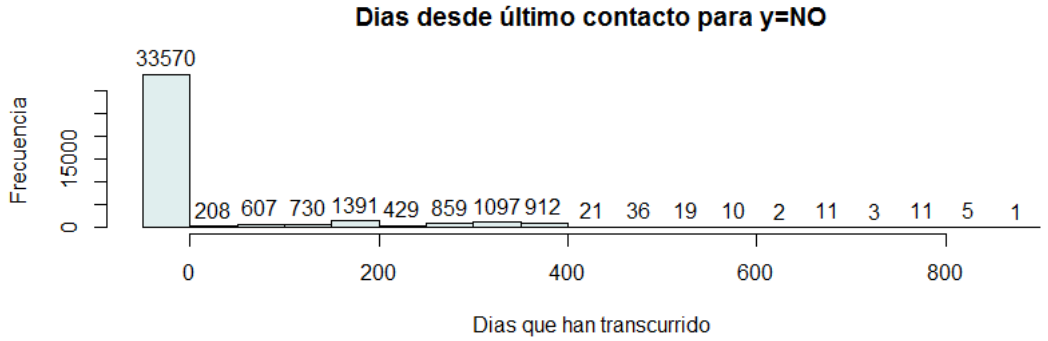
Depósito a Plazo	Sin Crédito de Vivienda	Con Crédito de Vivienda	Total
NO	16.727	23.195	39.922
SI	3.354	1.935	5.289
Total	20.081	25.130	45.211

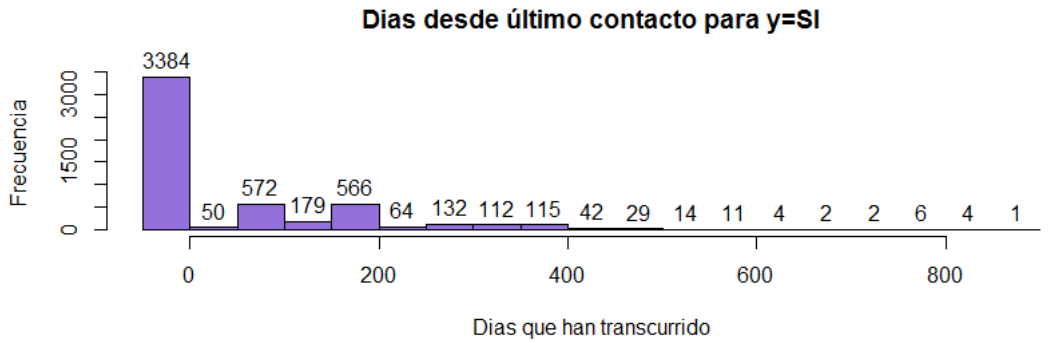
Análisis Descriptivo

Análisis Descriptivo

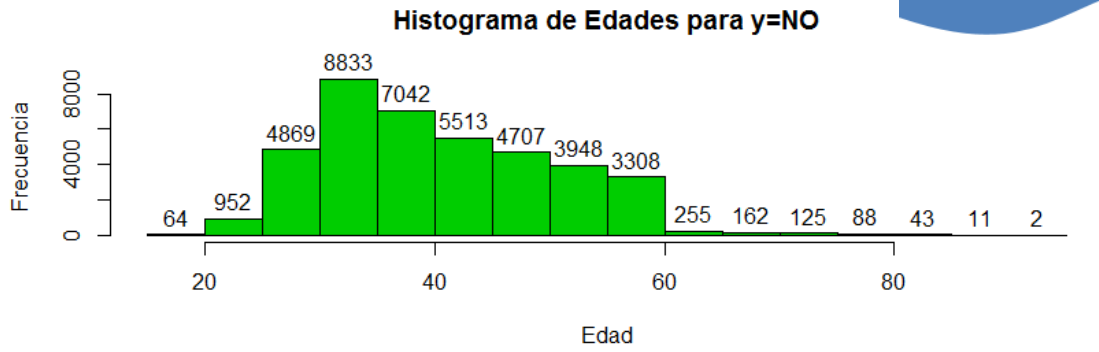
Variable pdays

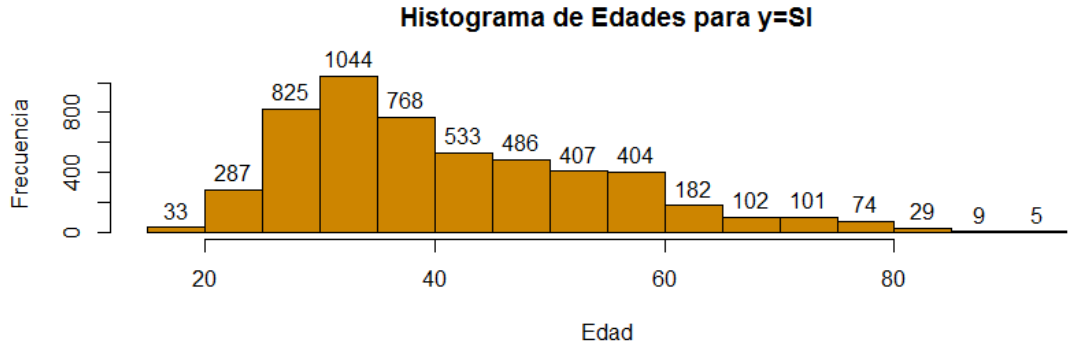
Días que han pasado desde el último contacto





Variable Edad





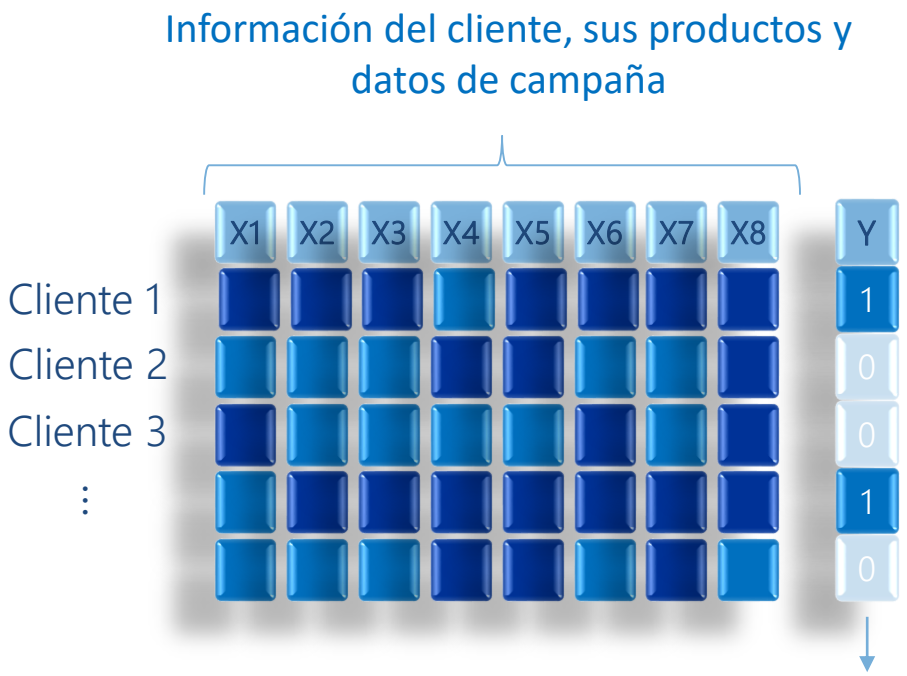
No se perciben mayores diferencias entre las distribuciones del SI y el NO

- Para otras variables no se observan patrones descriptivos que permitan explicar, qué caracteriza a los clientes que adquieren el producto de depósito o ahorro a plazo

Proceso de Modelado

Se requiere calcular la probabilidad de que un cliente se suscriba o no a un depósito a plazo (variable y), de acuerdo a la información disponible de sus datos demográficos, productos y la gestión de campañas de telemercadeo

Estructura de la Información



Respuesta binaria: Suscripción o NO

Observaciones sobre Los datos

- ✓ No hay presencia de NA's que puedan afectar las estimaciones
- ✓ Los datos proporcionados ya fueron pre-procesados y no muestran problemas de calidad que puedan afectar las estimaciones en los ejercicios de modelación
- ✓ **Los datos en análisis están desbalanceados** en relación a la variable y: De los 45.211 registros sólo 5.289 (11.7%) son suscripciones al depósito a plazo.

Proceso de Modelado

1. Base de modelado y base de prueba:

Se crea un subconjunto del 90% de los registros a través de un muestreo aleatorio como base de modelado y el restante 10% de los registros se usa para testeo y construir las métricas de clasificación.

2.Ejercicios de Modelación:

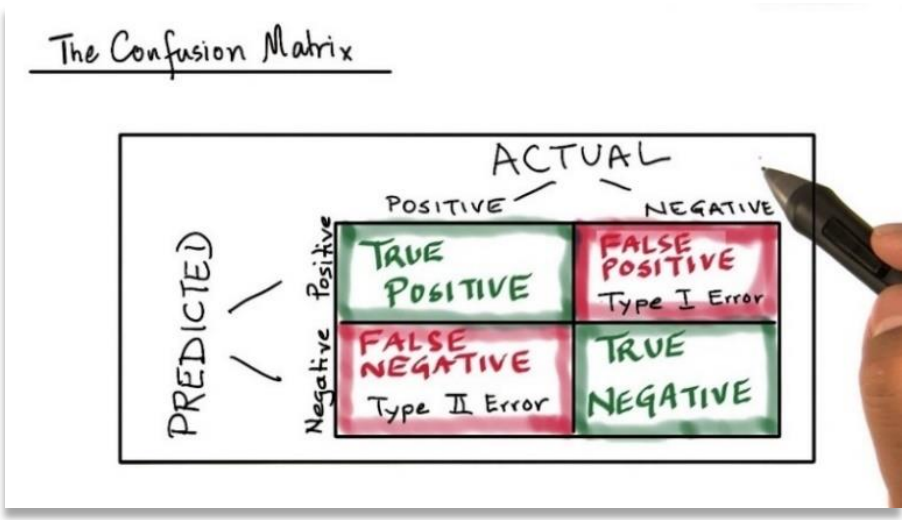
Modelos utilizados:

Regresión logística Múltiple y Decision Tree

Escenarios para construir los modelos:

Base de modelado (training), Over Sampling, Under sampling y ROSE (Random Over-Sampling Examples)

Evaluación de Modelos



- Se analiza la curva ROC y el AUC
- Matriz de Confusión:

Medidas de ajuste que se presentan como referencia son en base a la matriz de confusión:

	REFERENCIA	
PREDICCION	0/no	1/si
0/no	**A**	**B**
1/si	**C**	**D**

- *Sensitivity = A/(A+C)*
- *Specificity = D/(B+D)*
- *Prevalence = (A+C)/(A+B+C+D)*
- *PPV = (sensitivity * Prevalence)/((sensitivity*Prevalence) + ((1-specificity)*(1-Prevalence)))*
- *NPV = (specificity * (1-Prevalence))/(((1-sensitivity)*Prevalence) + ((specificity)*(1-Prevalence)))*
- *Detection Rate = A/(A+B+C+D)*
- *Detection Prevalence = (A+B)/(A+B+C+D)*
- *Balanced Accuracy = (Sensitivity+Specificity)/2*

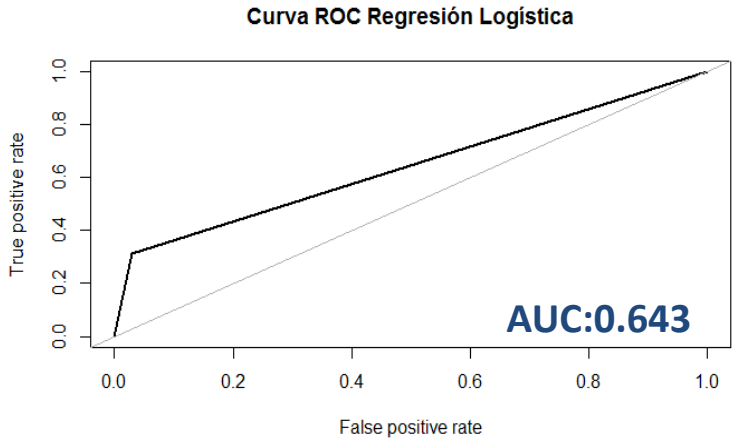
Métricas de clasificación con la Base: Training

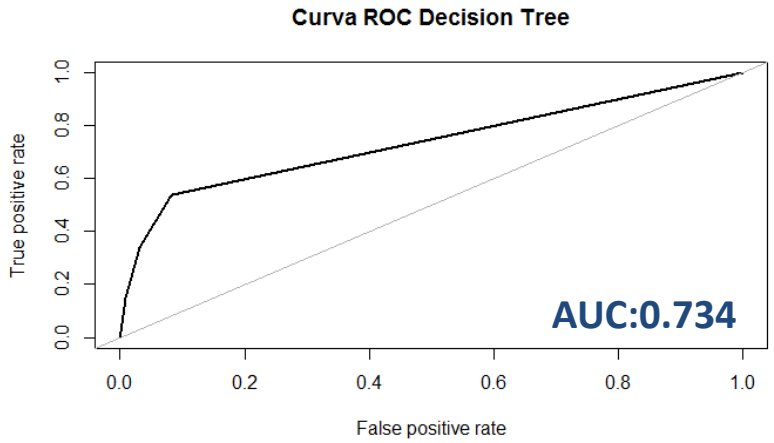
Multiple Logistic Regression

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3890  354
         1  115  163

               Accuracy : 0.8963
                 95% CI : (0.887, 0.905)
    No Information Rate : 0.8857
    P-Value [Acc > NIR] : 0.0124

                  Kappa : 0.3588
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9713
            Specificity : 0.3153
         Pos Pred Value : 0.9166
         Neg Pred Value : 0.5863
             Prevalence : 0.8857
         Detection Rate : 0.8602
   Detection Prevalence : 0.9385
      Balanced Accuracy : 0.6433
```

Decision Tree

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3878  341
         1  127  176

               Accuracy : 0.8965
                 95% CI : (0.8873, 0.9052)
    No Information Rate : 0.8857
    P-Value [Acc > NIR] : 0.01094

                  Kappa : 0.3766
 Mcnemar's Test P-Value : < 2e-16

            Sensitivity : 0.9683
            Specificity : 0.3404
         Pos Pred Value : 0.9192
         Neg Pred Value : 0.5809
             Prevalence : 0.8857
         Detection Rate : 0.8576
   Detection Prevalence : 0.9330
      Balanced Accuracy : 0.6544
```





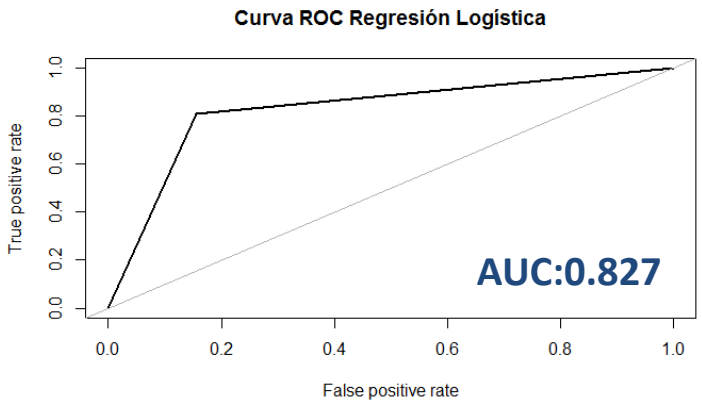
Métricas de clasificación con Over Sampling

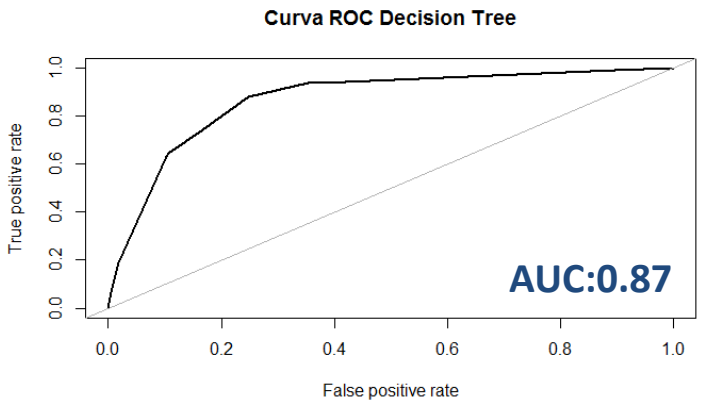
Multiple Logistic Regression

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3380   98
         1  625  419

               Accuracy : 0.8401
                 95% CI : (0.8291, 0.8507)
    No Information Rate : 0.8857
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4532
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8439
            Specificity : 0.8104
         Pos Pred Value : 0.9718
         Neg Pred Value : 0.4013
             Prevalence : 0.8857
         Detection Rate : 0.7475
   Detection Prevalence : 0.7691
      Balanced Accuracy : 0.8272
```

Decision Tree

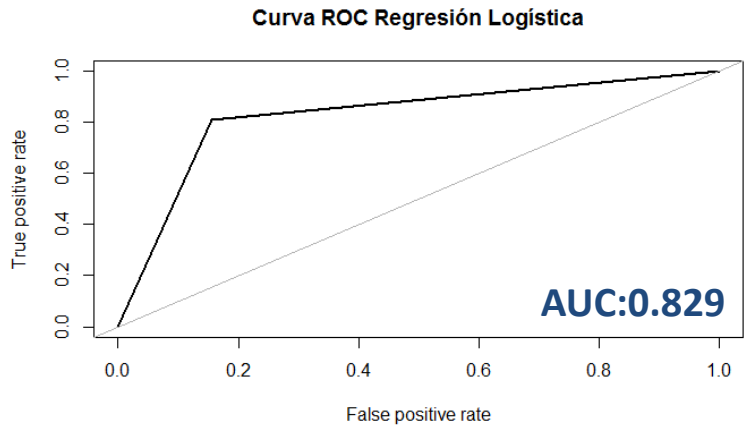
```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3010   61
         1  995  456

               Accuracy : 0.7665
                 95% CI : (0.7539, 0.7787)
    No Information Rate : 0.8857
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3546
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7516
            Specificity : 0.8820
         Pos Pred Value : 0.9801
         Neg Pred Value : 0.3143
             Prevalence : 0.8857
         Detection Rate : 0.6656
   Detection Prevalence : 0.6791
      Balanced Accuracy : 0.8168
```



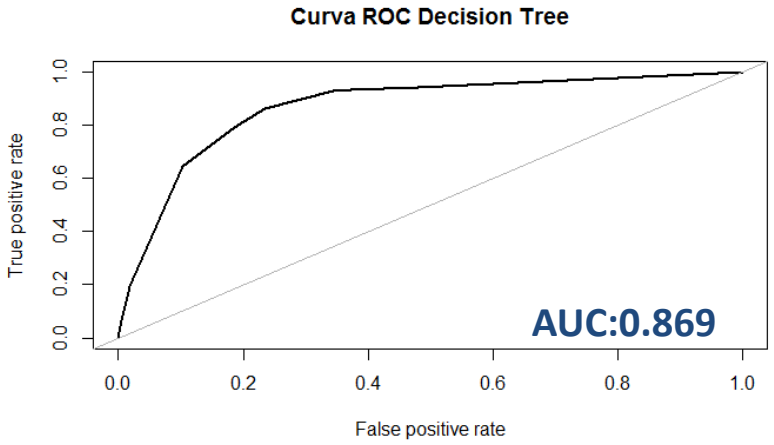


Métricas con Under Sampling

Multiple Logistic Regression

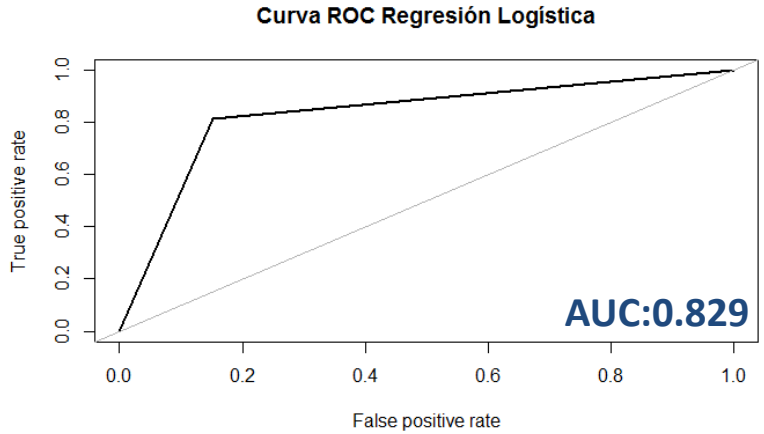


Decision Tree

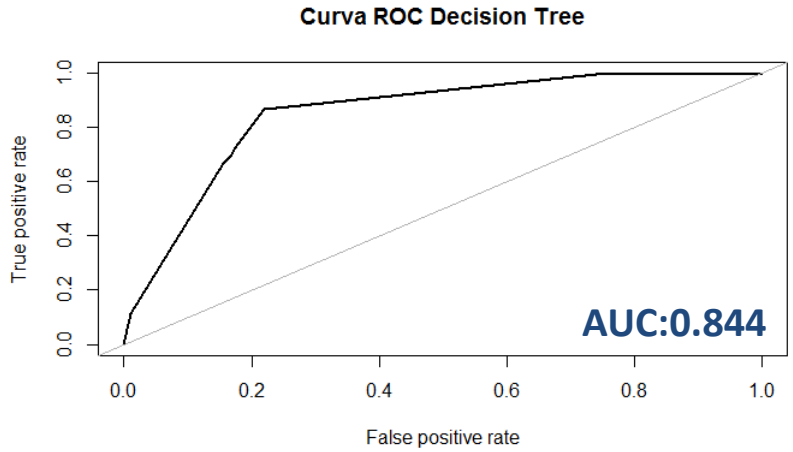


Métricas de clasificación con la Base: ROSE

Multiple Logistic Regression



Decision Tree



Observaciones sobre los distintos Modelos

1. Acciones sobre la muestra desbalanceada:

Independiente de la forma de muestrar para evitar sesgos en las estimaciones, se observa mejoría en las AUC con las distintas técnicas de análisis.

2. Cuál es el mejor modelo?

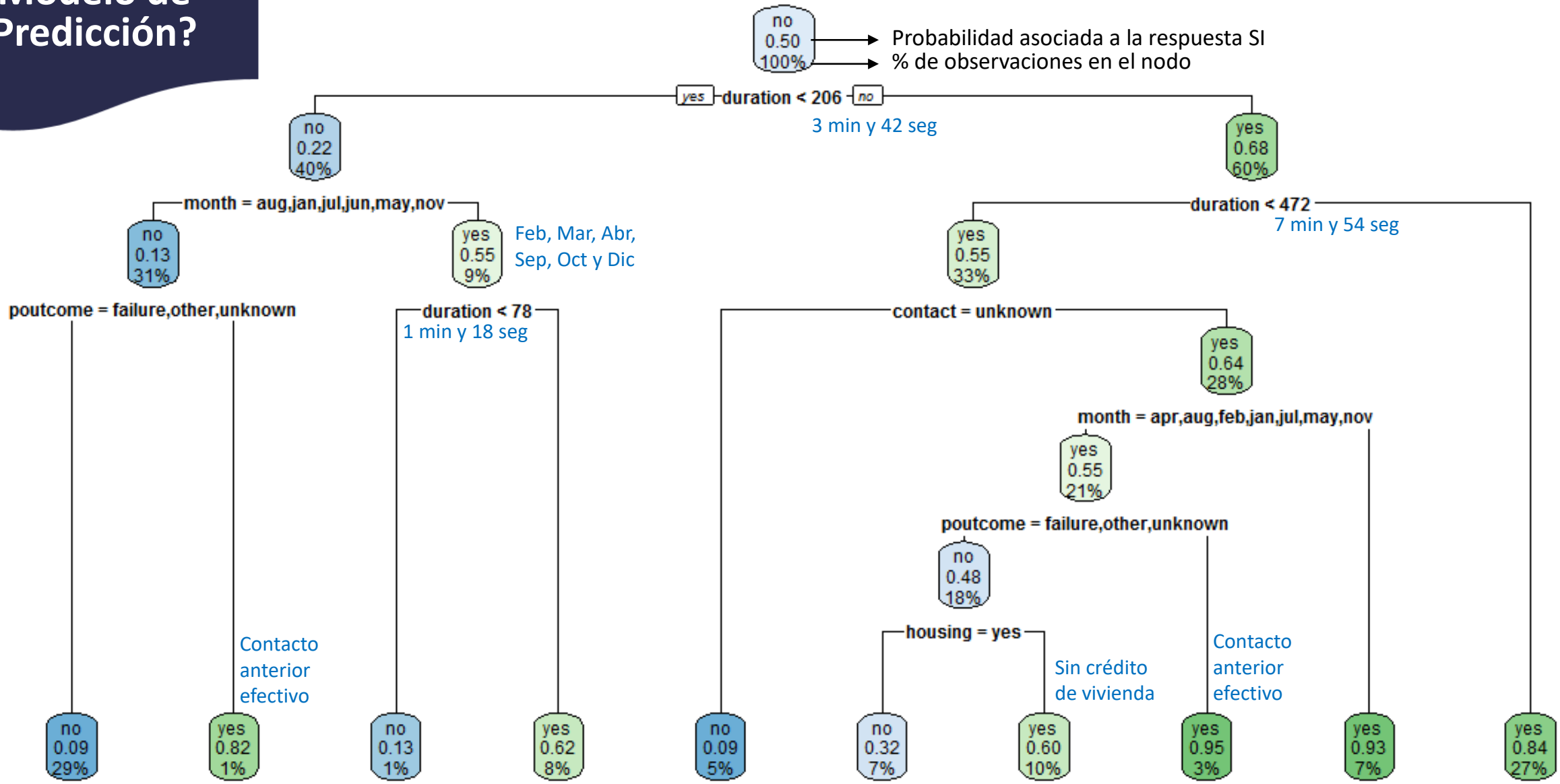
AUC Summary				
Technique/Data Base	**Training**	**Over Sampling**	**Under sampling**	**ROSE**
Multiple Logistic Regression	0,643	0,827	0,829	0,831
Decision Tree	0,734	0,87	0,869	0,85

Decision Tree produce las mejores AUC en los distintos ejercicios de modelación.

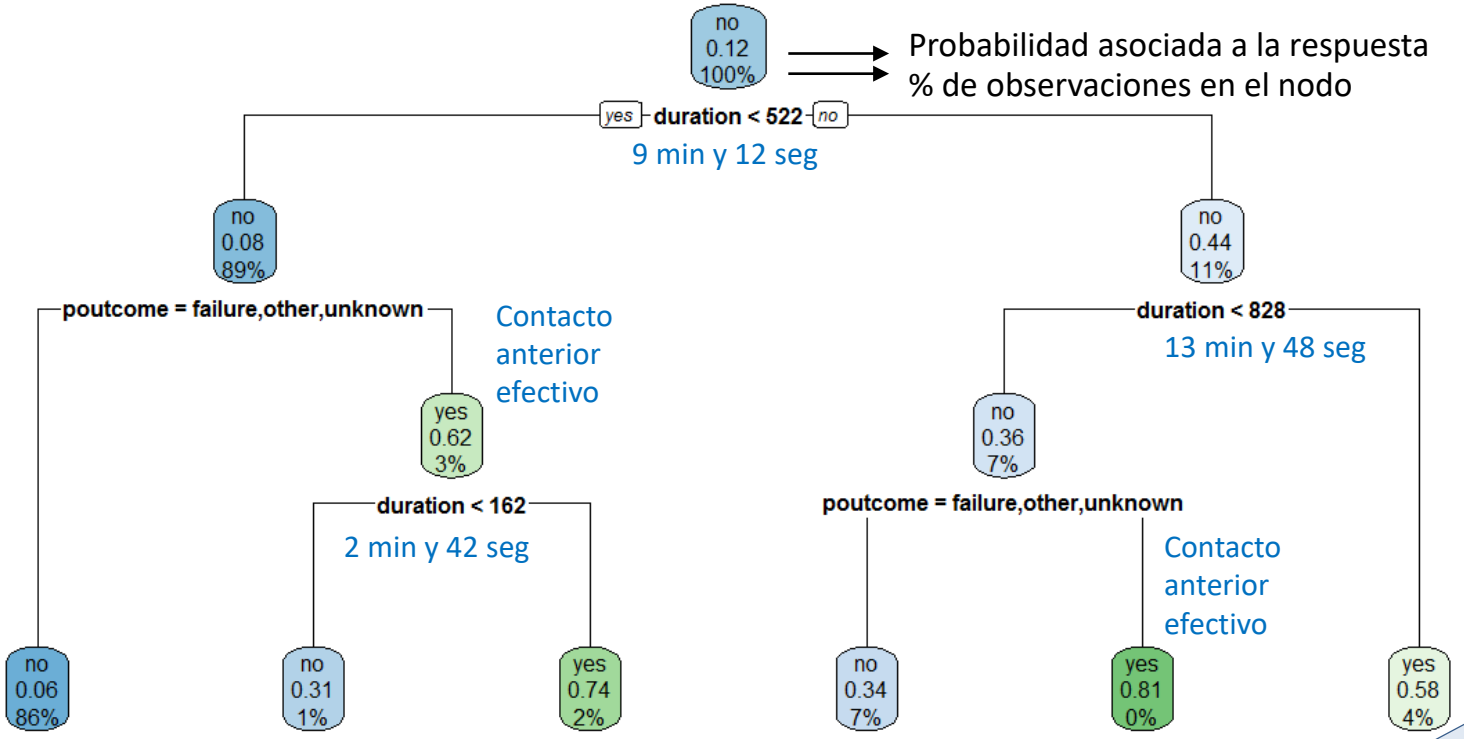
Cómo Interpretar el Modelo de Predicción?

Decision Tree

Gráfico con el escenario de Over Sampling

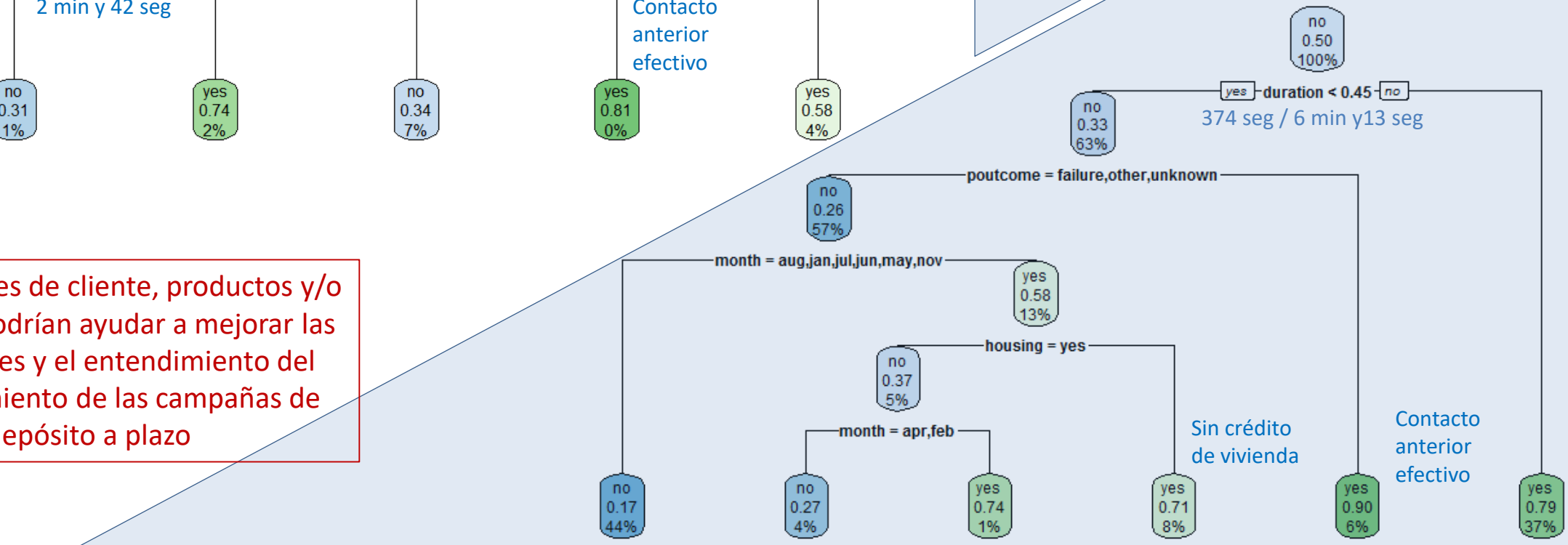


Decision Tree sin modificar datos y sin evaluar escenarios de predicción



Otras variables de cliente, productos y/o campañas podrían ayudar a mejorar las estimaciones y el entendimiento del comportamiento de las campañas de depósito a plazo

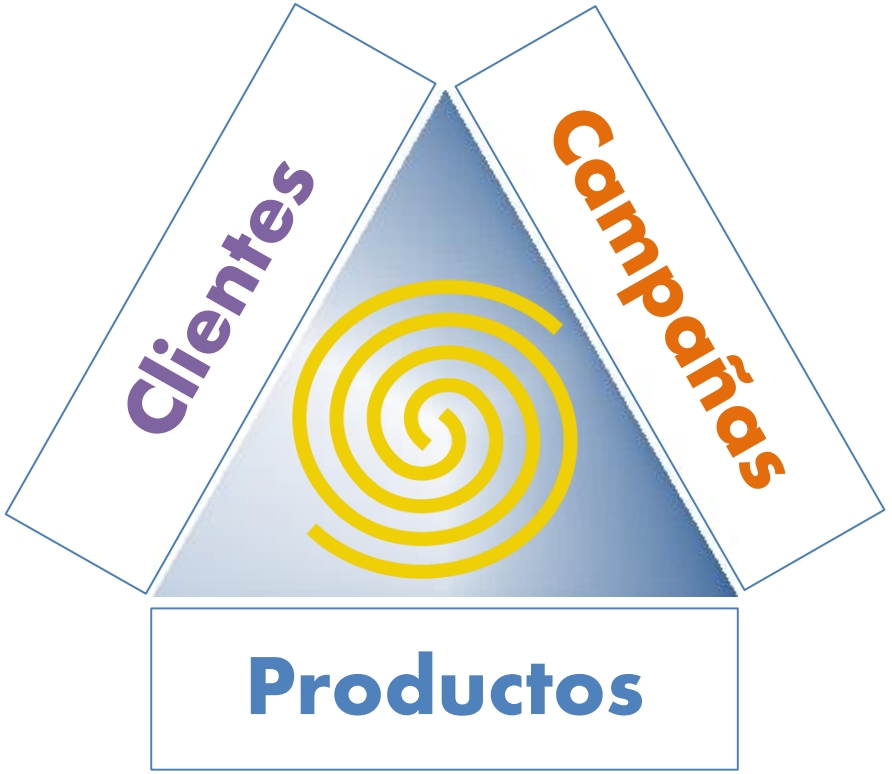
Decision Tree con variables numéricas estandarizadas y recategorización



Análisis Estratégico

3D del Territorio Estratégico de Análisis

Qué características demográficas influyen en la efectividad de las campañas y permiten que los clientes adquieran los productos?



Las campañas anteriores y los distintos contactos que se han tenido con el cliente, influyen en la decisión de adquirir el producto?

Los productos que tiene el cliente y su comportamiento de pago influyen en las campañas de depósito a plazo?

Características analizadas y principales hallazgos

- DATA IMPORTANTE PARA PREDECIR LA ADQUISICIÓN DEL PRODUCTO

✓ Deben explorarse más variables demográficas y/o posibles segmentaciones socioeconómicas o psicográficas

Edad
Ocupación
Estado civil
Educación

Tipo de contacto
- Mes del último contacto

Día del último contacto
- Duración de la última llamada

de contactos en la campaña vigente
Días desde el último contacto en campaña previa
de contactos en campaña previa
- Resultado de la campaña previa

✓ La posibilidad de que un cliente que no tiene crédito de vivienda adquiera el ahorro a plazo, es el doble frente a quienes si lo tienen (17% y 8%)

Balance anual
Incumplimiento en crédito
- Crédito de Vivienda

Crédito de Consumo

Sólo un 1% de los clientes analizados tenían incumplimiento en crédito

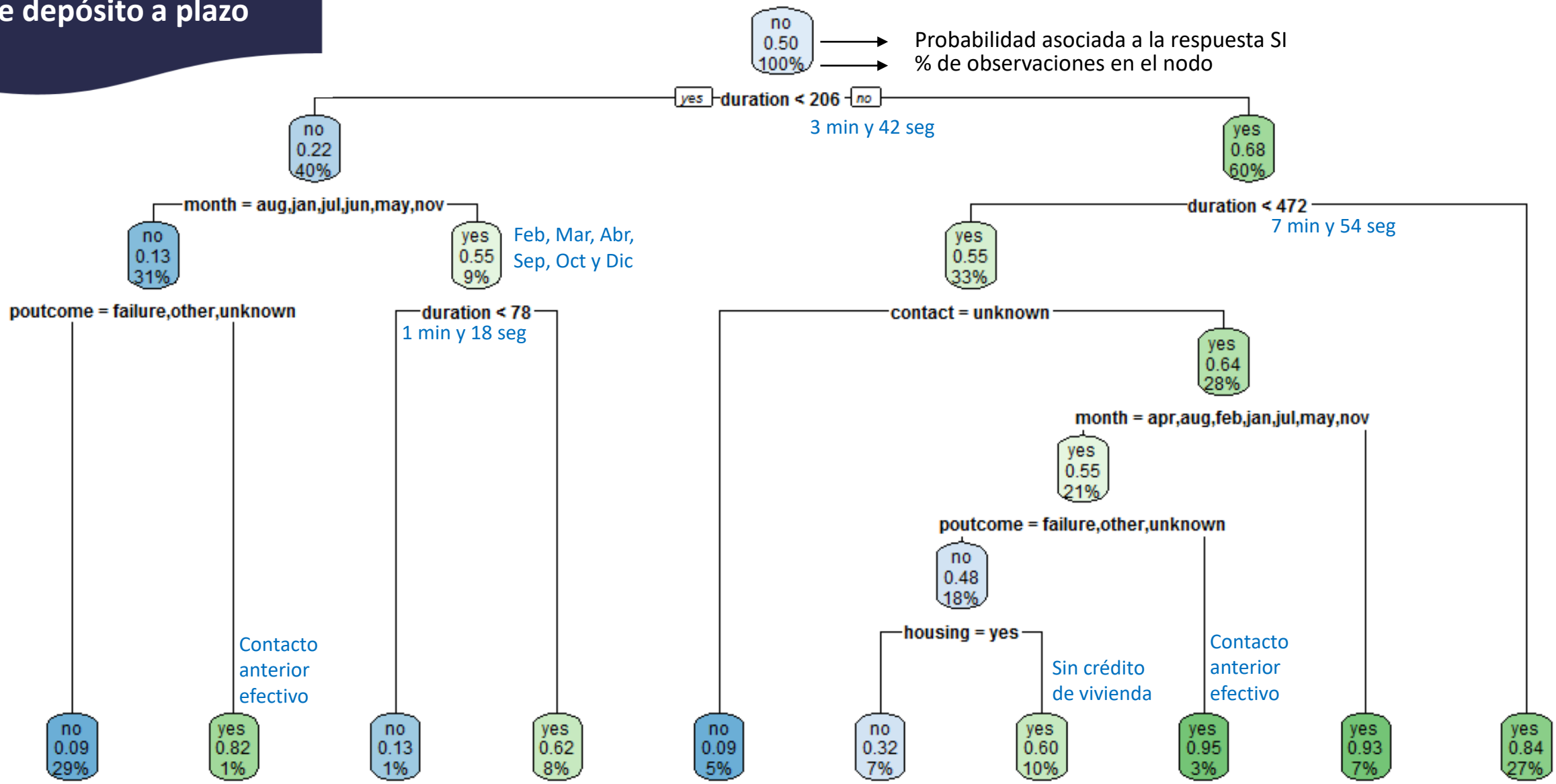
✓ Las llamadas de más de 3 minutos y medio suelen conducir a la adquisición del producto

✓ Los meses en los que la campaña de depósito a plazo suelen ser más efectivas son:

Feb, Mar, Abr, Sep, Oct y Dic

✓ Cuando no hay registro del resultado de la campaña anterior, es menos factible lograr la compra del producto

Decision Tree con datos de las campañas de depósito a plazo



Conclusiones y Recomendaciones

El análisis estadístico y los algoritmos de machine learning ayudan a la interpretación de información. Al usarlos en junto al conocimiento práctico de clientes, productos, y campañas, se busca la toma asertiva de decisiones de negocio

- ✓ El **performance de las campañas** se identifica como la dimensión que más afecta la adquisición del producto de depósito a plazo. Deben explorarse otros posibles factores de cliente (posibles segmentaciones) y de producto (transacciones y determinantes de la decisión de compra) para lograr mejores estimativos en los modelos de predicción explorados.
- ✓ Se requiere entender el propósito de las campañas previas realizadas por meses, para proponer cuáles deben ser las diferencias tácticas entre las campañas de: Feb, Mar, Abr, Sep, Oct y Dic y otros meses, donde se observa menor el chance de que un cliente adquiera el producto en la llamada de campaña.

Gracias

Referencias

Algunas de las web consultadas para desarrollar este ejercicio de modelación se listan a continuación:

http://apuntes-r.blogspot.com/2014/11/ejemplo-de-random-forest.html

https://www.researchgate.net/publication/275967793_Regresion_logistica_con_R

https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/

https://cran.r-project.org/web/packages/rpart/rpart.pdf

http://rischanlab.github.io/SVM.html

http://www.milbo.org/rpart-plot/prp.pdf

Anexo:

DataExploration.R - Código en R usado para obtener los resultados