UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL
DIPLOMADO DE INTELIGENCIA DE NEGOCIOS

MODELOS PREDICTIVOS PARA FUGA DE CLIENTES
DATA SCIENCE

Juan Pablo Corona Navarro
Esteban Mauricio Almuna Herrera
Bayron Damiani Retamal Gonzalez
Humberto Abel Gajardo Francino

Director:
Richard Weber

Semestre Otoño 2018

INDICE

I. INTRODUCCIÓN Y COMPRENSIÓN DEL NEGOCIO

i. Contexto

Diariamente, grandes volúmenes de datos son generados y procesados en las actividades operacionales de entidades tanto públicas como privadas. Por otro lado, las organizaciones deben tomar decisiones en relación a sus procesos de negocio, y ejecutar las acciones pertinentes. Lo ideal es que estas decisiones sean las correctas, para que los procesos entreguen resultados

La Inteligencia de Negocios (BI) entrega metodologías y recursos que permiten manipular los datos de una organización para la creación de conocimiento, el cual permite apoyar la toma de decisiones. Parte importante de la generación de conocimiento es la Minería de Datos, que se describe como el descubrimiento de patrones en los datos. Estos patrones son utilizados para análisis de clasificación y/o predicción para soportar los sistemas de toma de decisiones.

En este informe se presentan las etapas realizadas en la resolución de un caso de predicción de fuga de clientes en una entidad financiera. En particular, se describe el uso y análisis de distintas herramientas y metodologías en el campo de BI.

ii. Fuga de clientes

Esta tarea se centrará en el estudio del abandono voluntario y sus patrones de comportamiento de acuerdo a distintas variables el el modelo de datos.

Involuntario: Este tipo de fuga es fácil de identificar ya que existen variables o datos que apuntan a este abando tales como fallecimiento, morosidad, fraudes, etc.

Voluntario: Este tipo de fuga es más difícil de identificar ya que está dada por las decisiones que toma algún cliente, por reducción de ingreso, cancelación de productos, etc.

Entonces se puede decir que el abandono total (Cancelación de todos los contratos) es producido por el abandono real más el abandono técnico (Disminución de posiciones y actividad).

II. DESCRIPCIÓN DEL PROBLEMA

Una entidad financiera presenta altas tasas de fuga voluntaria. No hay una idea clara en relación al perfil de los clientes fugitivos ni de las causas que provocan estas fugas. Considerando esta situación se han encomendado dos tareas fundamentales. La primera tarea consiste en desarrollar un modelo predictivo que permita identificar de manera temprana a los clientes más propensos a fugarse. La segunda tarea consiste en la definición de políticas comerciales que permitan retener a los potenciales clientes fugitivos.

Como información adicional, en la siguiente tabla se muestran los beneficios/costos económicos de los aciertos y errores en las predicciones. Luego de aplicar el modelo de clasificación, la decisión comercial es definir a qué clientes cobrarles una tarifa normal y a qué clientes ofrecerles una oferta. La realización de una oferta corresponde a la acción comercial de retención. En caso de ser aplicada a un cliente no fugado tras el periodo de evaluación, se traduce en un costo de -100. Por otro lado, si la acción se aplica sobre un cliente con potencial fuga, se produce un beneficio de 1000. Por último, si no se aplica la acción a un cliente que posteriormente se fuga, se produce una pérdida de -5000

		Resultado	
		Fuga	No Fuga
Predicción	Fuga	1000	-100
	No Fuga	-5000	0

Considerando lo anterior, las preguntas expuestas por el gerente general son:

1. ¿Qué hacer para que estos clientes no se fuguen?
2. ¿Cuáles son las acciones comerciales que se deberían emprender con estos clientes?
3. ¿Han de ser aplicadas a todos los clientes?
4. ¿Cual es el costo de estas iniciativas?
5. ¿Qué tipo de herramienta tecnológica es necesaria?

Resultados esperados:

1. Desarrollo de un modelo predictivo de fugas que maximice el beneficio esperado.
2. Definir patrones característicos de clientes fugitivos y no fugitivos.
3. Sugerencia de al menos cinco acciones comerciales a emprender con los clientes que el modelo indique.
4. Predicción del conjunto VALIDACIÓN.

III. OBJETIVOS

i. Objetivos Generales

- Comprender las características del negocio y de los datos dentro del contexto del caso definido en esta tarea.
- Procesar los datos que serán utilizados para el modelo.
- Desarrollar un conjunto de modelos predictivos, analizarlos y seleccionar el más adecuado para solucionar el problema enunciado.
- Aplicar el modelo predictivo seleccionado al conjunto de datos a validar (clientes sin catalogar fuga/no fuga)
- Definir recomendaciones y medidas de acción sobre estos clientes una vez que han sido clasificados con el modelo predictivo.

ii. Objetivos Específicos

- Realizar un análisis exploratorio de los datos históricos que serán utilizados para el desarrollo del modelo supervisado.
- Implementar los procesos de imputación y limpieza sobre los datos históricos.
- Definir las transformaciones necesarias sobre los datos a usar en el desarrollo del modelo predictivo.
- Seleccionar las variables de acuerdo a su relevancia en relación a la variable objetivo.
- Construir distintos modelos predictivos con supervisión, mostrar detalles y parámetros de éstos y realizar un análisis comparativo en base a sus indicadores de performance y su interpretabilidad.
- Definir patrones de comportamiento de clientes fugados y no fugados.
- Tomar las imputaciones y transformaciones realizadas sobre los datos históricos y aplicarlas sobre los datos de validación.
- Utilizar el modelo predictivo seleccionado sobre los datos de validación para clasificar a los clientes que serán objetivo de la acción comercial para retenerlos.
- Analizar una serie de posibles políticas y acciones comerciales dirigidas a los actuales clientes de la entidad financiera.

IV. ANÁLISIS ESTADÍSTICO DE LOS DATOS

i. Descripción general

Para desarrollar el modelo predictivo solicitado, la entidad financiera ha entregado una base de datos histórica de sus clientes. Este conjunto de datos indica, para cada uno de los clientes, si éstos se han fugado o no.

En la siguiente tabla se detalla cada uno de los atributos de los clientes contenidos en la base de datos a usar en la implementación de los modelos. (dataset anonimizado con 20 variables)

Número	Variable	Descripción	Escala de medición
1	**ID**	Identificador del cliente	Escala Nominal
2	**Genero**	Genero del cliente	Escala Nominal
3	**Renta**	Renta en pesos	Escala de Razón
4	**Edad**	Edad en años	Escala de Razón
5	**NIV_Educ**	Nivel educacional	Escala de Ordinal
6	**E_Civil**	Estado civil	Escala Nominal
7	**COD_Ofi**	Código de la oficina	Escala Nominal
8	**COD_Com**	Código de la comuna	Escala Nominal
9	**Ciudad**	Ciudad de la oficina	Escala Nominal
10	**D_Marzo**	Deuda de Marzo	Escala de Intervalo
11	**D_Abril**	Deuda de Abril	Escala de Intervalo
12	**D_Mayo**	Deuda de Mayo	Escala de Intervalo
13	**D_Junio**	Deuda de Junio	Escala de Intervalo
14	**D_Julio**	Deuda de Julio	Escala de Intervalo
15	**D_Agosto**	Deuda de Agosto	Escala de Intervalo
16	**D_Septiembre**	Deuda de Septiembre	Escala de Intervalo
17	**M_Moroso**	Meses en Mora	Escala de Razón

18	**Monto**	Monto preaprobado	Escala de Intervalo
19	**Seguro**	Seguro de gravamen	Escala Nominal
20	**Fuga**	variable objetivo	Escala Nominal

ii. Lectura del Dataset

Se define el directorio de trabajo y se cargan las librerías correspondientes. Se lee el fichero de datos y se realiza una primera auditoría de datos.
Exploración de la estructura del Dataset

- Se carga las librerías correspondientes y se comienza con la carga del dataset llamado **"BASEFUGA.XLS"** y Conteo de missing values.

```
#data set
#Cargar BASEFUGA
#Definir carpeta de trabajo y abrir archivo BASEFUGA
#Session -> Set Working Directory -> To source file location
data_set<-read_excel("BASEFUGA.xls", sheet=1)

#Parte 1: AED
summary(data_set)

#Conteo missing values
na_count<-sapply(data_set,function(y)sum(length(which(is.na(y)))))
print(na_count<-data.frame(na_count))
```

Valores Nulos:

Se hace un conteo de valores nulos. El resumen de valores nulos es el siguiente:

Atributo	N/A Values	Atributo	N/A Values
ID	0	**D_Abril**	0
GENERO	11	**D_Mayo**	0
RENTA	0	**D_Junio**	0
EDAD	2	**D_Julio**	0
NIV_EDUC	11	**D_Agosto**	0
E_CIVIL	11	**D_Septiembre**	0
COD_OFI	0	**M_MOROSO**	0
COD_COM	3	**MONTO**	0
CIUDAD	3	**SEGURO**	0
D_Marzo	0		

Posteriormente, se procede a la detección de valores incorrectos y outliers (valores correctos pero que alejados de los límites del conjunto de valores del mismo tipo). Estos datos son marcados como nulos, para su posterior imputación.

Tomando en cuenta la cantidad de valores nulos en relación al total de registro de clientes, se realizará una imputación de datos mediante reemplazo por media/moda o funciones de relación con otras variables.

- Se continúa con la limpieza y se declaran las inconsistencias como valores perdidos.

```
data_set$EDAD[data_set$EDAD<0]<-NA
data_set$EDAD[data_set$EDAD>95]<-NA
#Codigo Oficina
data_set$COD_OFI[data_set$COD_OFI==0]<-NA
#NIV_EDUC
data_set$NIV_EDUC[data_set$NIV_EDUC=="EUN"]<-"UNV"

#Imputar Valores Perdidos por media/moda

#metodo para obtener moda
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
```

- Procede la imputación de datos sobre valores perdidos y obtenemos la moda.

```
#metodo para obtener moda
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

#GENERO
data_set$GENERO[is.na(data_set$GENERO)]<-getmode(data_set$GENERO)
#NIV_EDUC
data_set$NIV_EDUC[is.na(data_set$NIV_EDUC)]<-getmode(data_set$NIV_EDUC)
#E_CIVIL
data_set$E_CIVIL[is.na(data_set$E_CIVIL)]<-getmode(data_set$E_CIVIL)
#EDAD
data_set$EDAD[is.na(data_set$EDAD)]<-mean(data_set$EDAD,na.rm=TRUE)
#COD_OFI
data_set$COD_OFI[is.na(data_set$COD_OFI)]<-getmode(data_set$COD_OFI[data_set$CIUDAD=='VALPARAISO' & data_set$COD_COM==66])
#COD_COM
data_set$COD_COM[is.na(data_set$COD_COM) & data_set$COD_OFI==45]<-getmode(data_set$COD_COM[data_set$COD_OFI==45])
data_set$COD_COM[is.na(data_set$COD_COM) & data_set$COD_OFI==90]<-getmode(data_set$COD_COM[data_set$COD_OFI==90])
#CIUDAD
data_set$CIUDAD[is.na(data_set$CIUDAD) & data_set$COD_OFI==45 & data_set$COD_COM==88]<-getmode(data_set$CIUDAD[data_set$COD_OFI==45 & data_set$COD_COM==88])
data_set$CIUDAD[is.na(data_set$CIUDAD) & data_set$COD_OFI==90 & data_set$COD_COM==80]<-getmode(data_set$CIUDAD[data_set$COD_OFI==90 & data_set$COD_COM==80])
```

- Como se describen en las imágenes se aplicaron métodos en R para obtener data sin errores y posteriormente para su análisis y toma de decisiones. Se adjunta código R en anexo del informe.

iii. Transformaciones

Para que los modelos predictivos puedan utilizar los datos entregados, es necesario realizar algunas transformaciones sobre éstos. Las transformaciones usadas son: logaritmo(para variables monetarias) y variables dummy (para variables categóricas). La

siguiente tabla muestra un resumen de las transformaciones aplicadas sobre algunas variables independientes.

VAR	Original	Transformación	Var Final
MONTO	MONTO	ln_MONTO<-log(MONTO+1)	13, 14, 15, 16
GENERO	F: femenino M: masculino	Genero_bin 1-->F 0-->M	1 - 0
SEGURO	SEGURO_BIN	SEGURO_BIN: 1->SI, 0->NO	1 - 0
CIUDAD	SANTIAGO, ARICA, CONCEPCION, OTRAS	0->SANTIAGO, 1->ARICA, 2->CONCEPCION, 3->Otras	0 - 1 - 2 - 3
E_CIVIL	SOL,CAS,VIU, SEP	COD_E_CIVIL: 0->"SOL", 1->"CAS", 2->"VIU", 3->"SEP"	0 - 1 - 2 - 3
NIV_EDUC	BAS,MED,TEC,UNV	COD_NIV_EDUC: 0->"BAS" o "MED", 1->"TEC", 2->"UNV"	0 - 1 - 2
RENTA	RENTA	ln_RENTA<-log(RENTA + 1)	5.0, 7.5,10.0,12.5,15.0

iv. Selección de atributos.

Con las transformaciones y el escalamiento común aplicado a las variables independientes, se realiza una evaluación de relevancia de éstas, en relación a la variable objetivo(fuga/no fuga). Como primer paso, se genera la matriz de correlación entre variables (ver Anexo 1: Gráfico 1: Diagrama de correlación entre atributos). En este diagrama se observa que no hay ninguna correlación significativa entre todos los pares de atributos. En segundo paso, se utiliza la técnica del test Chi-Cuadrado. Este Test permite analizar la independencia de una variable respecto a otra a través del cálculo de la diferencia entre la distribución observada y la distribución teórica (tablas de contingencia), validando o refutando la hipótesis nula (las variables son independientes).

Se aplica el test a cada variable independiente, y se tienen los siguientes resultados: *Observación: Previamente se descartan las variables COD_OFI y COD_COM, debido a que CIUDAD toma relevancia como atributo geográfico.*

Variable	Grados de libertad	Chi-cuadrado	valor p
GENERO_BIN	1	516.34	<2.2E-16
EDAD	4	391.86	<2.2E-16
COD_NIV_EDUC	2	534.38	<2.2E-16
COD_E_CIVIL	3	201.32	<2.2E-16
COD_CIUDAD	3	770.96	<2.2E-16
M_MOROSO	3	345.43	<2.2E-16
SEGURO_BIN	**1**	**14.8**	**0.0001195**
RENTA	**4**	**9.9528**	**0.04123**
D_Marzo_Septiembre	4	30.033	4.82E-06
MONTO	4	40.573	3.30E-08

Se observa que para las variables SEGURO y RENTA, se tiene un valor chi-cuadrado significativamente bajo en comparación a las otras variables. Por otra parte, estos valores no permiten refutar la hipótesis nula. Por lo tanto, se decide descartar estas dos variables del conjunto de datos para el desarrollo de los modelos predictivos.

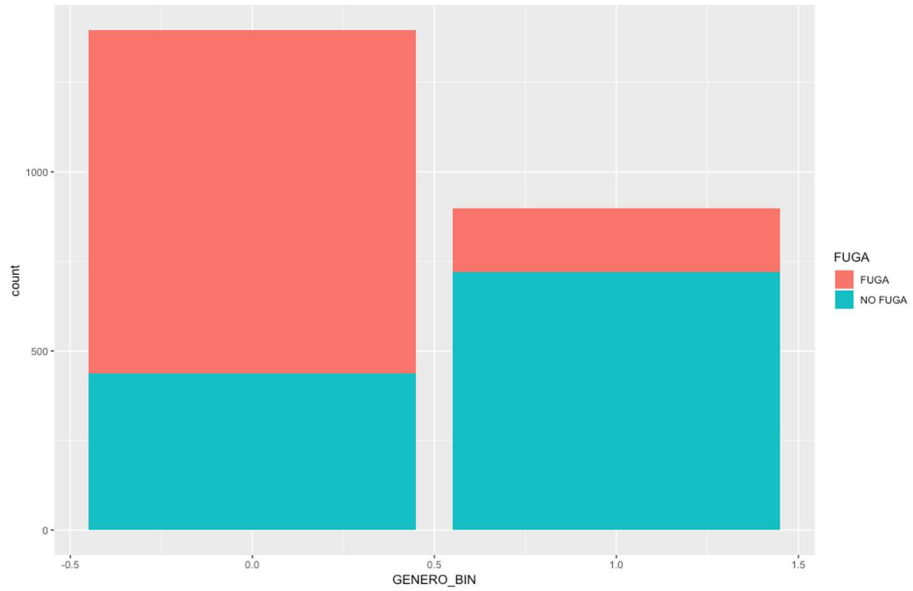
Con este proceso, se seleccionan las demás variables para la implementación de los modelos de clasificación binaria que se mencionan y describen en detalle en la siguiente sección de este informe.

V. Descripción de atributos

A continuación se hace una descripción de los atributos seleccionados para la implementación de los modelos predictivos.

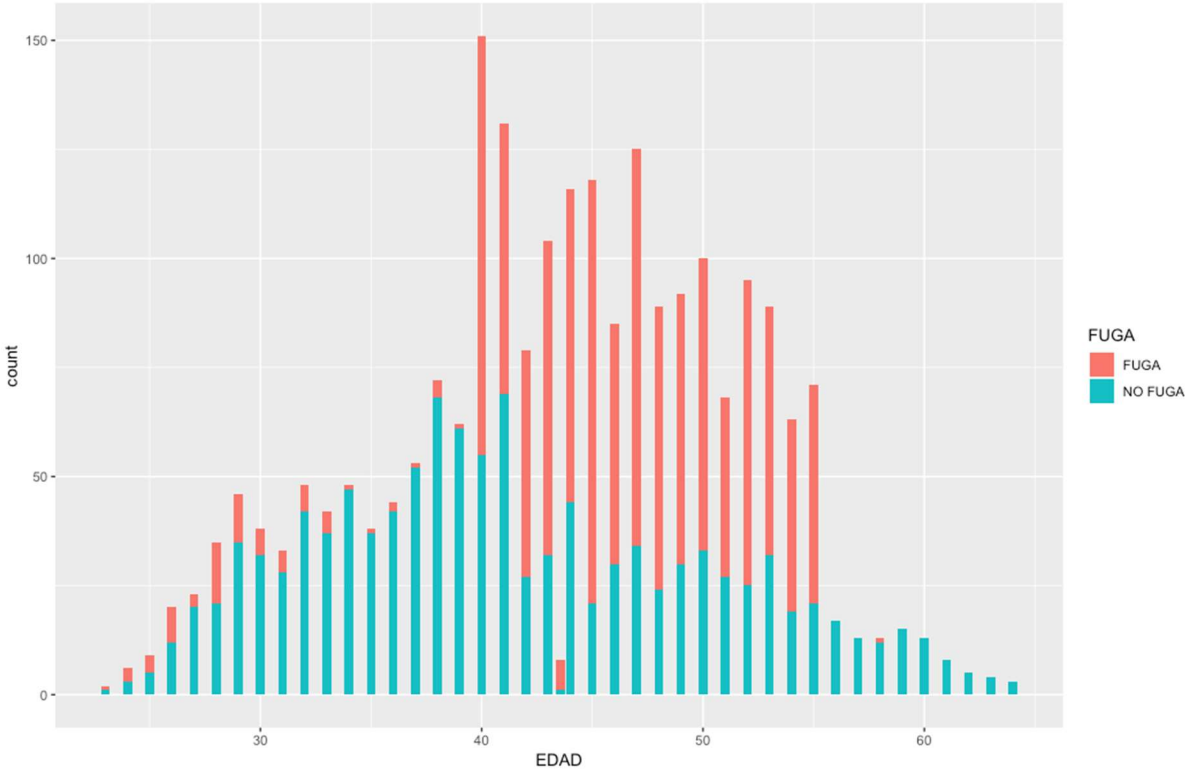
GENERO_BIN

Se nota con gran exactitud que el género femenino si bien es menor al género masculino, el porcentaje de fuga es menor en mujeres y mucho mayor en hombres



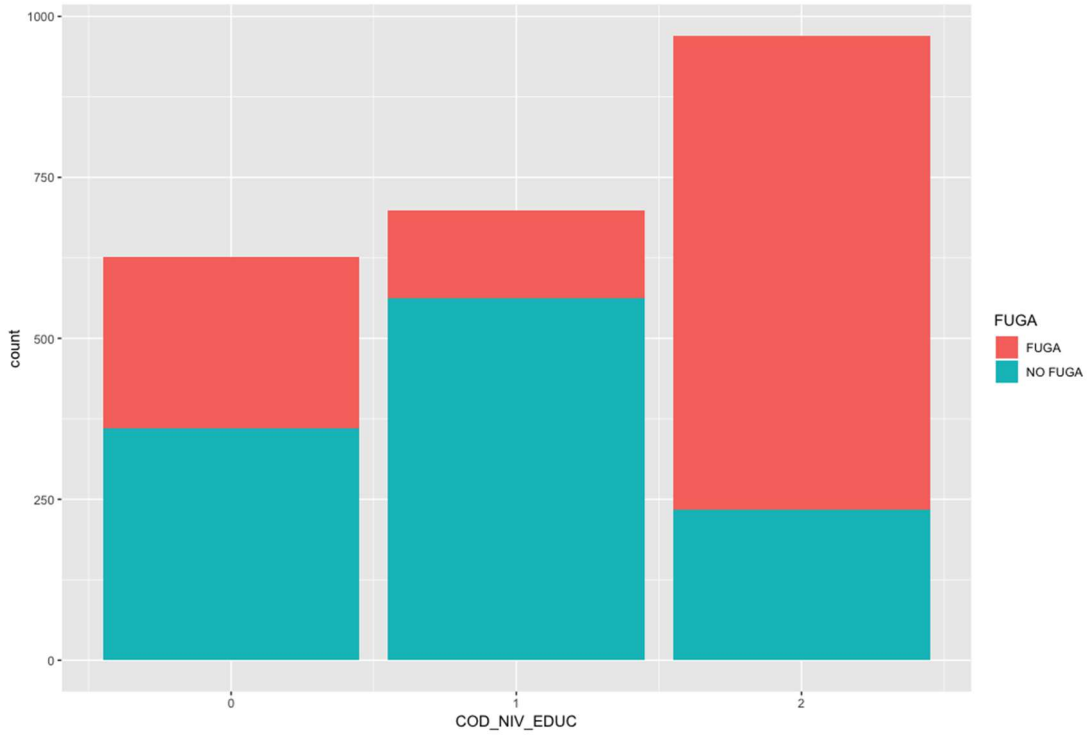
EDAD

Se observa que los clientes menores a 40 años presentan porcentajes mayores en no-fuga: Desde los 40 hasta los 55 se observan porcentajes altos en fuga. Para 55 años o más se aprecia un mayor porcentaje en no-fuga.



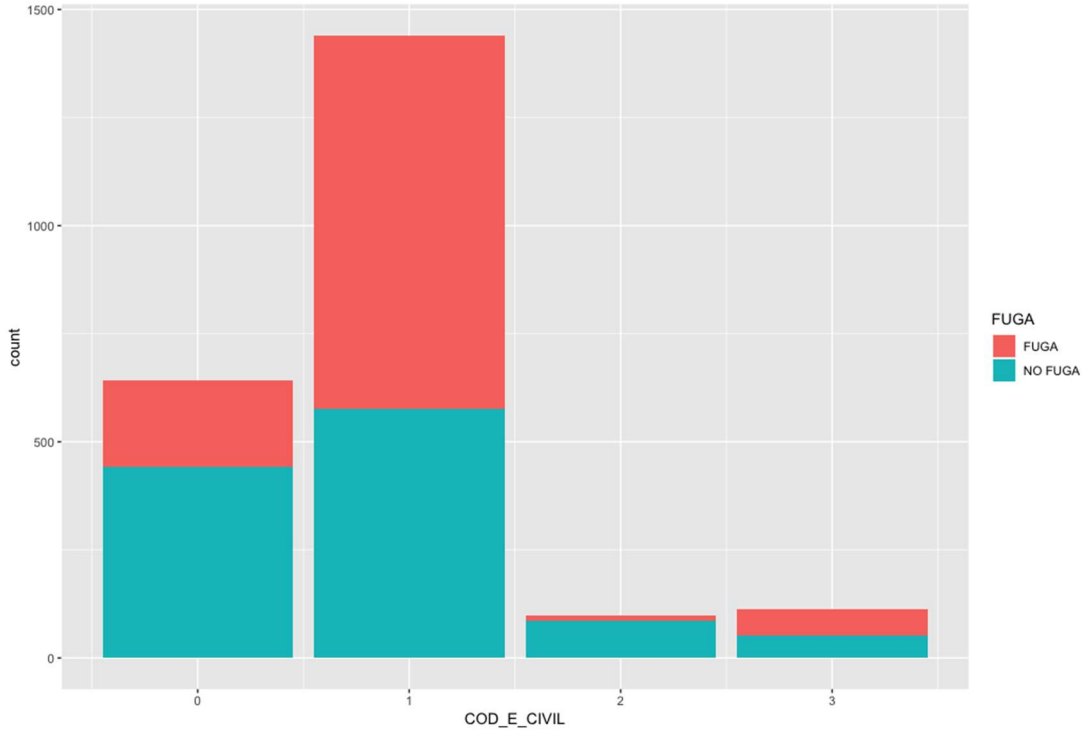
COD_NIV_EDUC

Se aprecia que existe gran fuga de clientes universitarios y menos fuga de clientes el los niveles técnicos y básicos, aunque se debe destacar que también existe fuga en nivel básico, pero no en la frecuencia alta que presentan los universitarios.



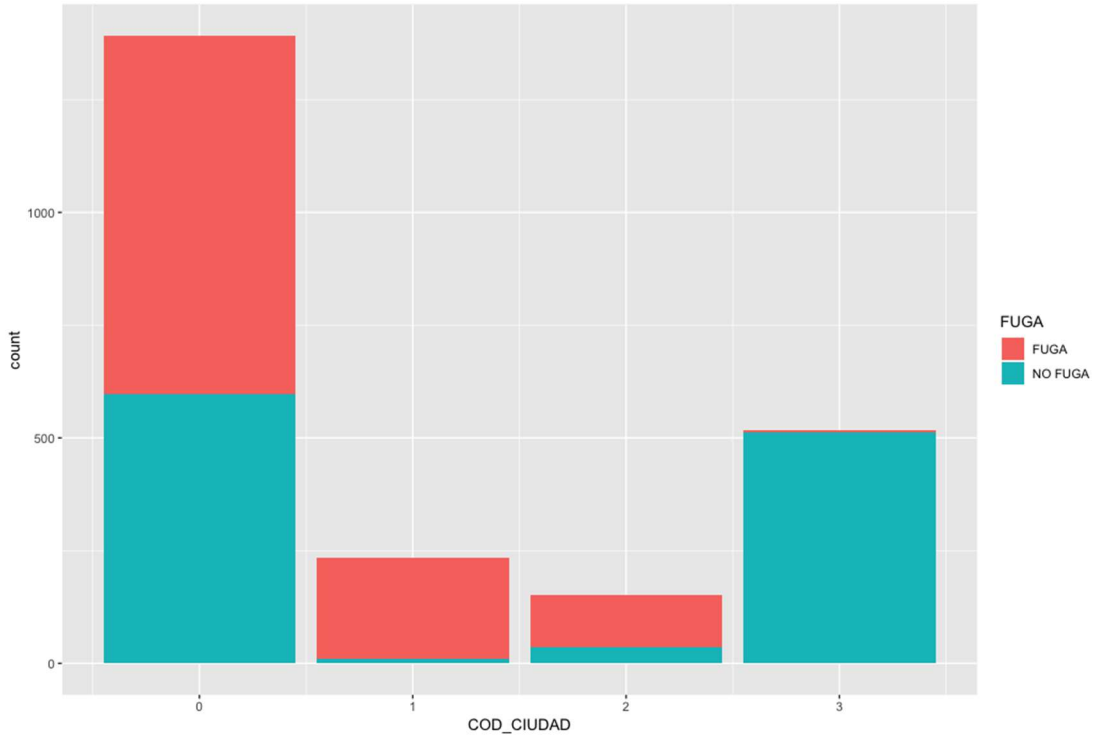
COD_E_CIVIL

Se logra detectar gran cantidad de clientes fugados en los clientes que son casados, luego en menor cantidad de frecuencia los solteros, sin embargo existe gran cantidad de clientes no fugados en solteros en relación a los fugados del mismo estado civil.



COD_CIUDAD

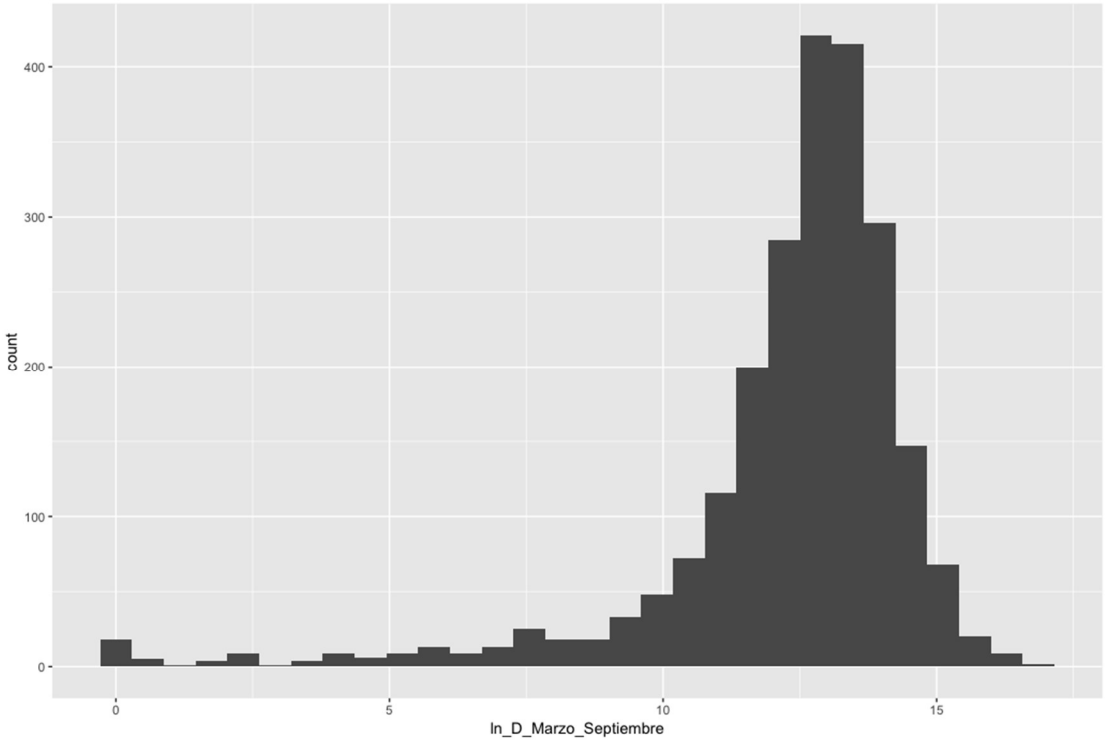
Se observa que existe mayor fuga en la ciudad de Santiago, aunque también se logra detectar una gran cantidad de clientes no fugados en la misma ciudad, sin embargo en Arica existe la mayor cantidad de fuga de clientes sin mantener un equilibrio. En regiones la fuga de clientes es muy baja, por no decir que casi no existe fuga, son casos muy puntuales.



ln_D_Marzo_Septiembre

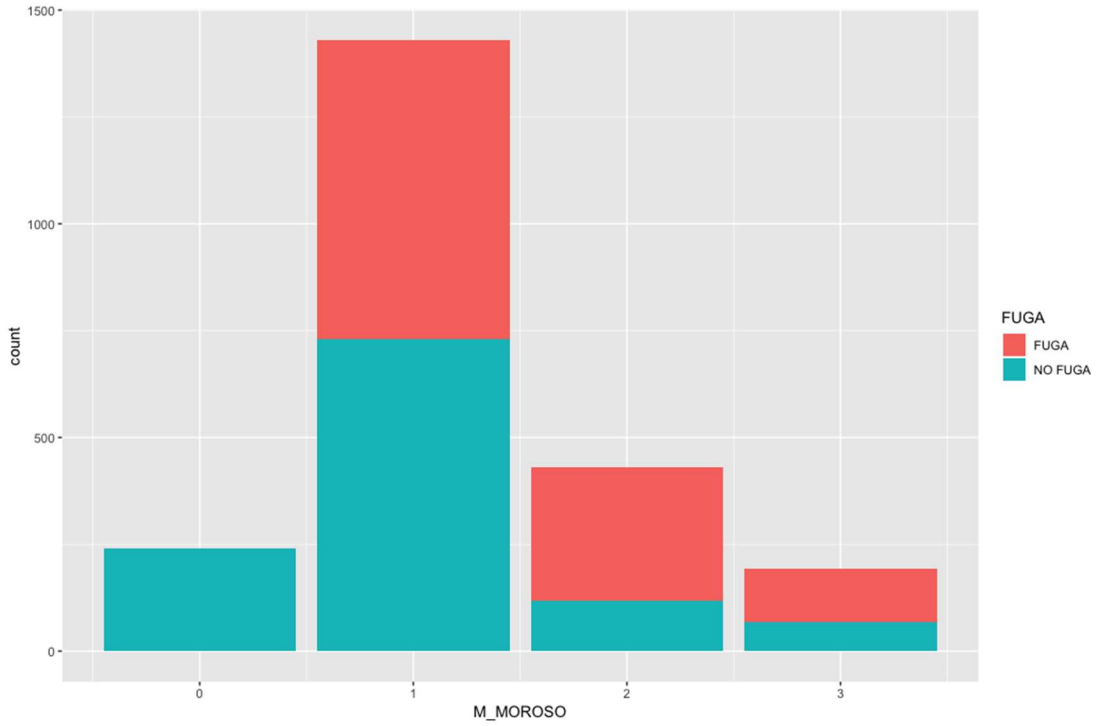
Se detecta que la frecuencia de deuda al aplicar logaritmo en el mes de septiembre mantiene un promedio entre 10 y 15 en relación a la frecuencia del gráfico.

Se logra notar que entre 0 y 5 hay una variación muy baja.



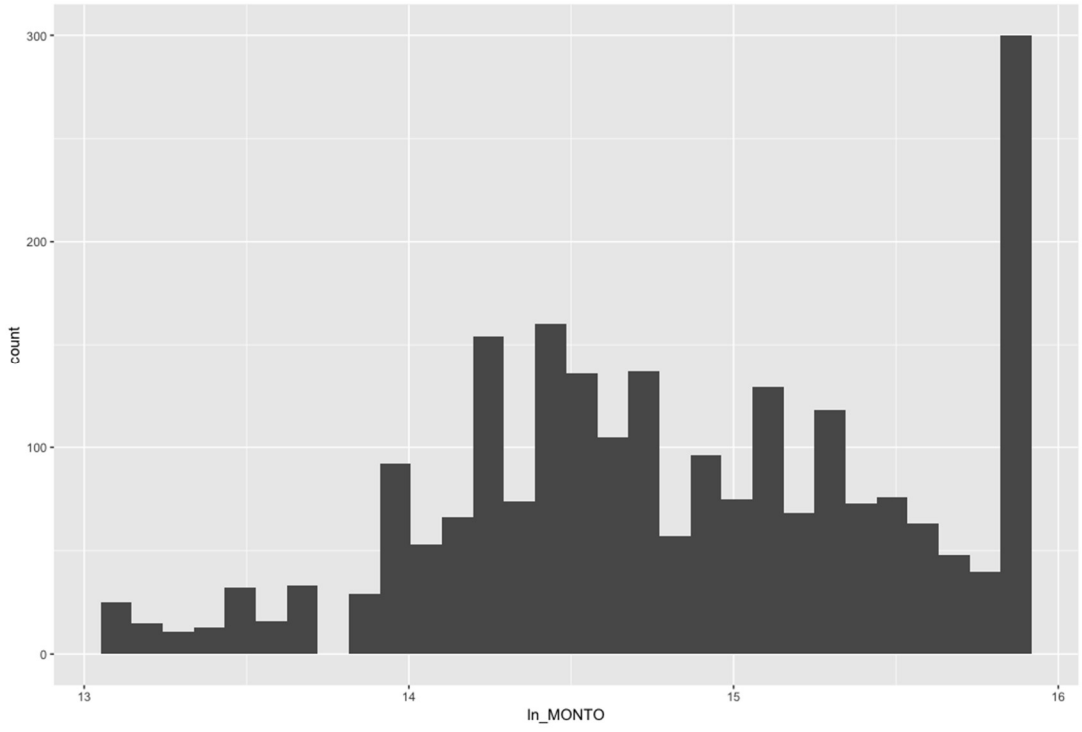
M_MOROSO

En este gráfica se logra apreciar que existe mora de 1 mes frecuentemente para los clientes que se fugan y le sigue la fuga a los clientes que presentan 2 meses de mora.



ln_MONTO

Se verifica que al aplicar logaritmo al monto de preaprobación ln_MONTO<-log(MONTO+1) en relación a los clientes en su frecuencia se observa un aumento en la escala cercana al 16 según el gráfico.



V. MODELOS DE MINERÍA DE DATOS

i. Regresión Logística

Descripción: La Regresión Logística es un modelo de regresión que permite predecir el valor de una variable categórica en base a las variables independientes. Modela la probabilidad de un evento(variable a predecir) como función en base a otros factores(atributos) a través de una función logística. Por otro lado, este modelo de tipo "wrapper" permite observar la relevancia de cada variable independiente. Esto es de utilidad en dos puntos: primero, en la selección de atributos (variables poco relevantes pueden ser descartadas para el modelo) y segundo, en la definición de patrones.

Detalles:

Residuos de desviación:

Min	*1Q*	*Mediana*	*3Q*	*Max*
-3.3748	*-0.6192*	*0.0911*	*0.5744*	*2.9986*

Coeficientes:

	Estimado	*Error Std*	*valor z*	*Pr(>|z|)*	
(Intercept)	*2.5399*	*0.4107*	*6.184*	*6.25e-10*	***
GENERO_BIN	*2.1142*	*0.1264*	*16.719*	*< 2e-16*	***
EDAD	*-3.9140*	*0.3174*	*-12.332*	*< 2e-16*	***
COD_NIV_EDUC	*-1.4081*	*0.1414*	*-9.960*	*< 2e-16*	***

COD_E_CIVIL	*-0.8853*	*0.2599*	*-3.406*	*0.000659*	***
COD_CIUDAD	*2.2932*	*0.1604*	*14.296*	*< 2e-16*	***
ln_D_Marzo_Septiembre	*-1.1166*	*0.4511*	*-2.475*	*0.013322*	*
M_MOROSO	*-2.1601*	*0.2580*	*-8.373*	*< 2e-16*	***
ln_MONTO	*1.3414*	*0.2378*	*5.640*	*1.70e-08*	***

---
*Códigos de Significancia: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Resultados:

2294 muestras
8 predictores
2 clases: 'FUGA', 'NO_FUGA'

Sin pre-procesamiento
Remuestreo: Validación Cruzada (10 fold)
Resumen de tamaños de muestras: 2064, 2065, 2065, 2064, 2064, 2064, ...
Resultados de Remuestreo:
ROC: 0.892818 Sens:0.8399317 Spec:0.8011769

ii. K-Nearest Neighbors

Descripción: K vecinos más cercanos es un método de clasificación supervisada que determina la clase de un elemento en base a una función de densidad asociada a los elementos ya clasificados más “cercanos”(esta “cercanía” se calcula usando la distancia euclidiana). Un objeto es clasificado con la clase de mayor densidad dentro de los K vecinos más cercanos.

Resultados:

2294 muestras
8 predictores
2 clases: 'FUGA', 'NO_FUGA'

Sin pre-procesamiento
Remuestreo: Validación Cruzada (10 fold)
Resumen de tamaños de muestras: 2064, 2065, 2065, 2064, 2064, 2064, ...
Resultados de Remuestreo:
ROC: 0.9445635 Sens:0.9147415 Spec:0.8625337

ROC fue usado para seleccionar el modelo óptimo usando el valor más grande.

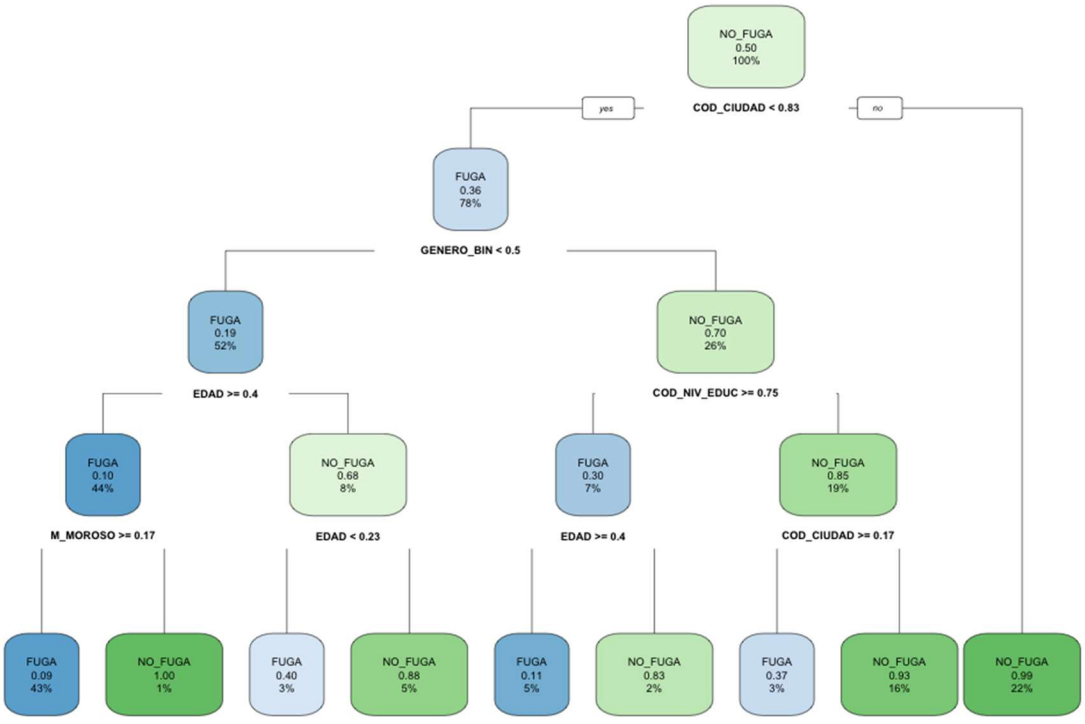
The valor final usado por el modelo fuel was k = 9.

Para ver el valor ROC con distintos valores de K (cantidad de vecinos cercanos), ver Anexo 1, Gráfico 2.

iii. Árbol de Decisión

Descripción: Es un modelo de predicción que utiliza una serie de construcciones lógicas que sirven para representar una serie de condiciones sucesivas para completar el proceso de clasificación. Cada nodo del árbol es una decisión basada en una condición sobre un atributo. Sobre esa condición, se hace una clasificación preliminar. Cada subconjunto resultante de esa clasificación pasa por una secuencia de nodos de decisión hasta generar subconjuntos con alta densidad en cierta categoría. Este modelo

Detalles: En el siguiente diagrama se observa el árbol de decisión resultante del desarrollo del modelo con los atributos seleccionados.



Resultados: Se utilizó como parámetro de complejidad(cp) el valor 0.009, el cual entregó la sensibilidad máxima.

2294 muestras

8 predictores
2 clases: 'FUGA', 'NO_FUGA'

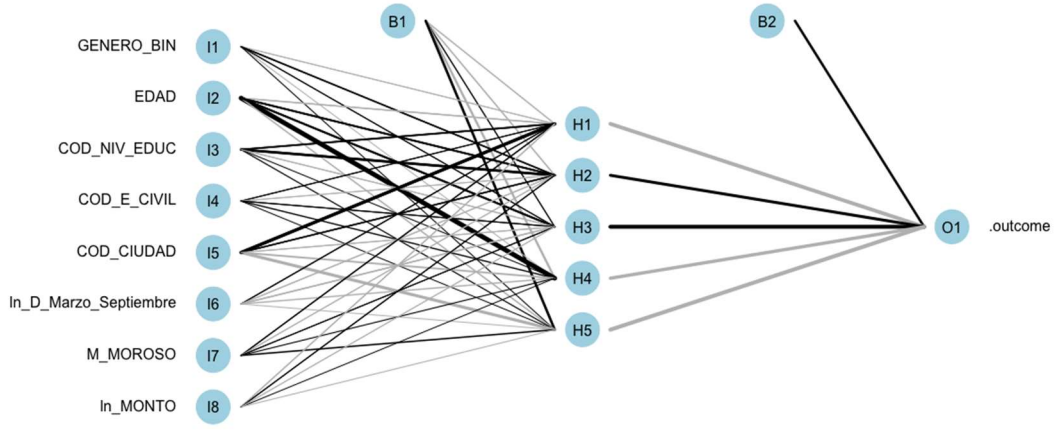
Sin pre-procesamiento
Remuestreo: Validación Cruzada (10 fold)
Resumen de tamaños de muestras: 2064, 2065, 2065, 2064, 2064, 2064, ...
Resultados de Remuestreo:

ROC	*Sens*	*Spec*
0.9409628	*0.9446126*	*0.872901*

iv. Red Neuronal

Descripción: En el contexto de clasificación supervisada, Una Red Neuronal Artificial es un modelo computacional que recibe como entrada una representación de las variables independientes de una entidad a clasificar y entrega como salida la clasificación asignada, de acuerdo al aprendizaje adquirido al recibir los datos utilizados para su desarrollo. Una red neuronal está compuesta por nodos o "neuronas", de entrada, de capa escondida y de salida, y de conexiones entre éstas, las cuales tienen asignadas un "peso" o medida de relevancia de conexión.

Detalles: En el siguiente diagrama se muestra la red neuronal desarrollada a partir de las variables independientes seleccionas



Resultados:

2294 muestras

8 predictores
2 clases: 'FUGA', 'NO_FUGA'

Sin pre-procesamiento
Remuestreo: Validación Cruzada (10 fold)
Resumen de tamaños de muestras: 2064, 2065, 2065, 2064, 2064, 2064, ...
Resultados de Remuestreo:
ROC: 0.9586383 Sens:0.9332247 Spec:0.8738156

ROC fue usado para seleccionar el modelo óptimo usando el valor más grande.
Los valores finales usados para el modelo fueron size = 5 y decay = 0.1.

Para observar los valores ROC con distintas cantidades de nodos escondidos(size) y distintos valores en peso de decaimiento(decay), ver Anexo 1, Gráfico 3.

v. Support Vector Machines

Descripción: Support Vector Machines es un algoritmo de aprendizaje supervisado que consiste en la construcción de uno o más hiperplanos de separación, mediante un vector entre dos puntos, en un espacio de dimensionalidad muy alto, que es utilizado para la separación entre clases. Para este caso, se utilizó la SVM-Kernel.

Resultados:

2294 muestras
8 predictores
2 classs: 'FUGA', 'NO_FUGA'

Sin pre-procesamiento
Remuestreo: Validación Cruzada (10 fold)
Resumen de tamaños de muestras: 2064, 2065, 2065, 2064, 2064, 2064, ...
Resultados de Remuestreo:
ROC: 0.9597112 Sens:0.9068157 Spec:0.8867616

ROC fue usado para seleccionar el modelo óptimo usando el valor más grande.
Los valores finales usados para el modelo fueron sigma(desviación típica) = 0.1010907 y C = 1.

Para observar los valores ROC para distintos valores de C (peso de cada observación), ver Anexo 1, Gráfico 4.

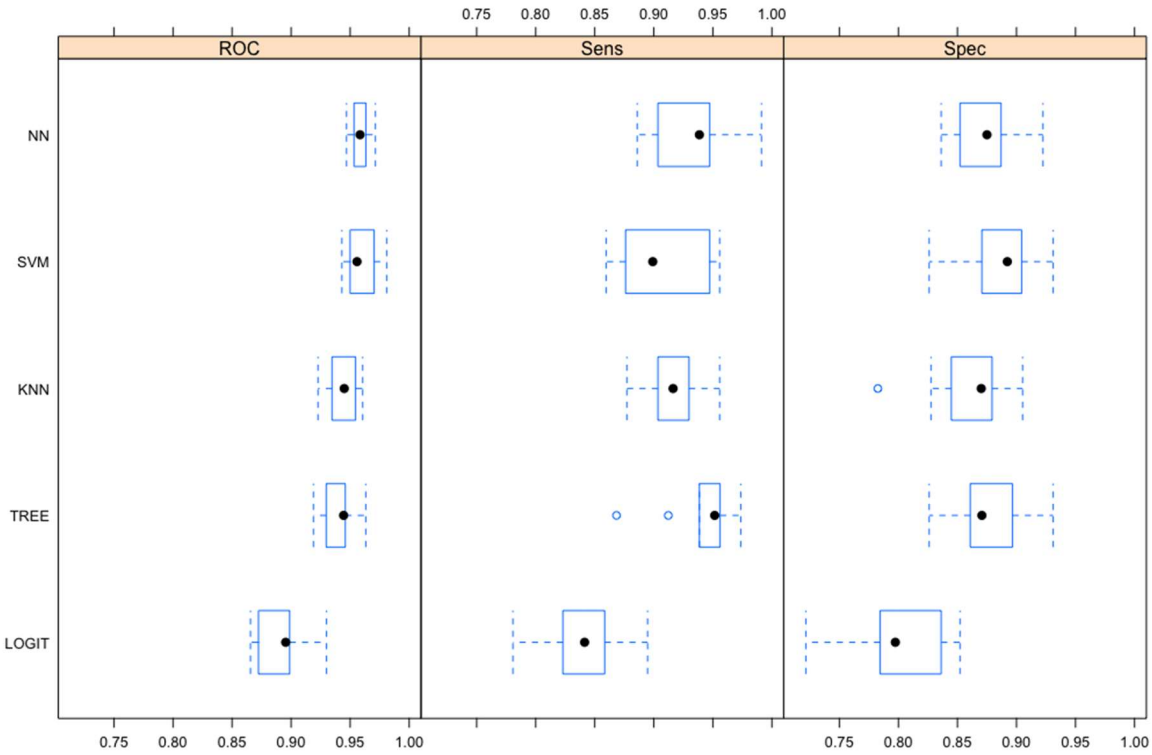
VI. SELECCIÓN DEL MODELO

- **Árbol de decisión**

Una vez que se han desarrollado los modelos predictivos, se procede con su análisis y se realiza la selección del modelo más adecuado según este análisis.

Se ha decidido considerar dos criterios: indicadores de performance e interpretabilidad.

Performance: Para cada modelo, se obtienen los siguientes valores de performance: Curva ROC (Razón de verdaderos positivos versus falsos positivos), Sensibilidad (razón de verdaderos positivos) y Especificidad (razón de verdaderos negativos). Con estos valores, se tiene el siguiente gráfico con diagramas de caja (box-plot) para cada modelo(Red Neuronal: NN, Support Vector Machines: SVM, K-Nearest Neighbors: KNN, Árbol de Decisión: TREE y Regresión Logística: LOGIT), para cada indicador:



Como se puede observar, todos los modelos obtienen una mediana del valor ROC mayor a 0.89, por lo que se puede concluir que, en general todos los modelos son buenos (valor ROC más cercano a 1 que a 0.5). En cuanto a la sensibilidad y la especificidad, no hay diferencia significativa entre NN, SVM, KNN y TREE. Sin embargo en LOGIT se ven valores menores en relación a los otros modelos mencionados anteriormente.

Viendo la matriz de costos (ver "Descripción del problema"), los falsos negativos (Predicción: No fuga, Resultado: Fuga) generan el mayor costo, seguido de los falsos positivos (Predicción: Fuga, Resultado: No fuga). En este sentido, la sensibilidad del modelo toma mayor relevancia, por sobre la especificidad.

Interpretabilidad: Como parte de la tarea asignada, se pide definir patrones de conducta en clientes fugados y no fugados. En este punto, es recomendable seleccionar un modelo que, dentro de sus características entregue información en relación a la relevancia de los atributos utilizados en la categorización aplicada. Los modelos NN, SVM y KNN son considerados como "caja negra", es decir, no entregan información relacionada a la importancia de cada uno de los atributos utilizados en su implementación. En cambio, los modelos TREE y LOGIT brindan información relacionada con la relevancia o peso de atributos.

Si bien los modelos NN, SVM y KNN obtienen los valores ROC más altos, se decidió descartarlos de la selección debido a que no permiten interpretabilidad. Por lo tanto, el modelo seleccionado para la predicción de fuga es **Árbol de Decisión** (TREE). Este modelo obtiene buenos valores de ROC, Sensibilidad y Especificidad, y, cuenta con un alto grado de

interpretabilidad, ya que realiza una secuencia de clasificaciones de acuerdo a los atributos más relevantes.

VII. POLÍTICAS COMERCIALES

i. Patrones

Al registrar una cantidad significativa de clientes que se han fugado voluntariamente, y, con los resultados de la predicción, se sugiere aplicar un conjunto de políticas comerciales, con el propósito de retener a los clientes, en especial a los clientes clasificados como posibles fugas.

Como fase preliminar, se aplica la clasificación de los clientes registrados en la base VALIDACION utilizando el modelo predictivo (ver adjunto). Este conjunto de clientes clasificados será utilizado para validar el modelo posterior a la aplicación y evaluación de las políticas comerciales, mediante la comparación entre la predicción y el resultado final (fuga/no fuga) para los clientes del conjunto de validación.

Para definir las posibles políticas comerciales y los clientes objetivos de éstas, es necesario definir algunos patrones en relación al comportamiento de los clientes, a través de sus atributos. Observando el Árbol de Decisión desarrollado en el modelo, en particular, los nodos de decisión y las clasificaciones generadas a partir de éstos, se tienen algunos tipos de comportamiento que pueden ser considerados patrones. Algunos patrones detectados son:

- Clientes en ciudades distintas a Santiago, Arica y Concepción (las tres con mayor cantidad de clientes), tienen altísima probabilidad de no fuga (99%)
- Clientes en Santiago, Arica o Concepción, tienen probabilidad de fuga de 36%
- Clientes del género femenino, que vienen en Santiago, Arica o Concepción, con nivel educacional distinto a "Universitario", tienen alta probabilidad de no fuga (85%)
- Clientes del género masculino entre 32 y 39 años, tienen alta probabilidad de no fuga (88%)
- Clientes del género masculino menores a 39 años sin morosidad, tienen 100% no fuga.
- Clientes del género masculino menores a 32 años, tienen probabilidad de fuga de (40%)

Lo primero que de detecta es la relevancia de la ciudad en donde se ubican las sucursales. Se observa una diferencia en el comportamiento de fuga entre los clientes de Santiago/Arica/Concepción y el resto de las ciudades. En ese contexto, podría ser interesante analizar la situación en todas las ciudades donde la entidad financiera tiene sucursales en dos aspectos: primero, la gestión en el servicio a los clientes, y segundo, en la competencia (nivel de oferta). En este segundo aspecto, las políticas comerciales juegan un rol fundamental, como factores que distinguen a la entidad financiera por sobre sus competidoras. También se observa la relevancia del nivel educacional en el caso de los clientes de género femenino, y la relevancia de la edad y los meses de morosidad en los clientes de género masculino. Todo esto permite definir más adelante las características de los clientes que serán parte de la aplicación de la o las políticas comerciales.

ii. Acciones comerciales

Considerando las observaciones descritas anteriormente, se realiza la definición de políticas comerciales. Para establecer estrategias comerciales se crearon campañas interrelacionadas que simulan un ciclo de cliente en un plazo de 3 meses, que abarca 2 de los principales atributos a la hora de definir si un cliente tiene mayor probabilidad de fuga.

Se definieron 5 acciones comerciales que se concentra en un plazo de 3 meses, donde del total de los potenciales fugados abarca el 50% de los clientes con mayor margen a nuestra empresa.

La decisión de tomar exclusivamente al 50% de los clientes, se debe a analizar el margen de los clientes actualmente fugados y descubrir que aproximadamente la mitad de ellos no genera márgenes positivos para la empresa.

Para simplificar el ejercicio se hará el supuesto de que al extrapolar lo anterior a la predicción de fuga actual, el comportamiento y margen aportado a la empresa del total de los potencialmente fugados se comporta de la misma manera.

El modelo definió que la probabilidad de que un cliente sea fugado que corresponda en las regiones de Santiago, Concepción y Arica disminuirá con un marketing estratégico de fidelización.

Basado en lo anterior se declaró que los 2 objetivos comerciales para fidelizar a los clientes se buscaron 5 acciones serán:

- Fidelizar a los clientes al aumento de tarjetas de crédito, consiguiendo que cada cliente tenga al menos 2 tarjetas de crédito.
- Fidelizar al Incrementar el crédito a través del uso de las tarjetas de crédito de nuestros clientes.

Antes de implementar las campañas comerciales se definió el perfil del cliente estándar demográficamente hablando, concluyendo que, del total de clientes potencialmente fugados, un 84% son hombres y un 16% mujeres, que la concentración de mayor fuga se encuentra entre clientes de 40 y 55 años y que la concentración de se encuentra en las regiones de Santiago, Concepción y Arica.

Definiendo estos atributos comunes se decidió realizar 2 campañas el primer mes, 3 campañas el segundo y 1 el último mes.

- **Campaña 1: Tarjeta adicional para cliente o familiar como adicional.**

La campaña tendrá como objetivo **incrementar el número de tarjetas** de crédito por cada cliente para posteriormente aumentar su uso.

Para ello se asumirán los siguientes supuestos:

1. Contar con los datos de contacto del total del grupo objetivo.
2. La evaluación de riesgo de los clientes permite 1 tarjeta adicional para cada uno de ellos.

El target de la campaña abarca a todos los clientes potencialmente fugados que tengan sólo 1 tarjeta de crédito, y el enganche de venta para los clientes será el beneficio de una tarjeta adicional sin costo de mantención por 3 meses.

Se comunicará a los clientes por distintos canales como SMS, e-mail, llamadas de call center, utilizando la multicanalidad como una oportunidad en la efectividad de la campaña.

- **Campaña 2: Convenios en comercios asociados**

La campaña tendrá como objetivo **Incrementar el crédito** a través del uso de las tarjetas de crédito que actualmente posee. Para ello se asumirá el siguientes supuestos: Se contará con los datos de contacto del total del grupo objetivo.

El target de la campaña abarca a todos los clientes potencialmente fugados que no tengan uso en sus tarjetas de crédito, y el enganche de venta para los clientes serán distintos descuentos a través de alianzas estratégicas con restaurantes, tiendas de deportes, tecnología y vestuario.

Se comunicará a los clientes por distintos canales como SMS, e-mail, llamadas de call center, utilizando la multicanalidad como una oportunidad en la efectividad de la campaña.

SEGUNDO MES: El segundo mes se realizarán campañas de fidelización para el aumento del uso de tarjeta, descuento en la primera compra al obtener una tarjeta adicional y se repetirá la campaña de convenios asociado al uso de tarjetas del primer mes.

- **Campaña 3: Plan de fidelización con acumulación de puntos**

La campaña tendrá como objetivo **Incrementar el crédito** a través del uso de las tarjetas de crédito que actualmente posee.

Para ello se asumirán los siguientes supuestos:

1. Contar con los datos de contacto del total del grupo objetivo.
2. Ha utilizado la tarjeta previamente y aún cuenta con capacidad de endeudamiento.
3. Existe actualmente un catálogo de puntos en la compañía.

El target de la campaña abarca a todos los clientes potencialmente fugados que tengan uso en sus tarjetas de crédito el mes anterior, y el enganche de venta para los clientes serán $25.000 de regalo para canjear cualquier artículo del catálogo de fidelización de la empresa, por cada $350.000 en compras con sus tarjetas de crédito.

Se comunicará a los clientes por distintos canales como SMS, e-mail, llamadas de call center, utilizando la multicanalidad como una oportunidad en la efectividad de la campaña.

- **Campaña 4: Devolución primera compra**

La campaña tendrá como objetivo **incrementar el número de tarjetas** de crédito por cada cliente para posteriormente aumentar su uso.

Para ello se asumirán los siguientes supuestos:

1. Contar con los datos de contacto del total del grupo objetivo.
2. La evaluación de riesgo de los clientes permite 1 tarjeta adicional para cada uno de ellos.

El target de la campaña abarca a todos los clientes potencialmente fugados que en la campaña del mes anterior no hayan incrementado el número de sus tarjetas, y el enganche de venta para los clientes será la devolución de $10.000 en su primera compra.

Se comunicará a los clientes por distintos canales como SMS, e-mail, llamadas de call center, utilizando la multicanalidad como una oportunidad en la efectividad de la campaña.

- **Campaña 5: Extensión de campaña convenios en comercios asociados**

La campaña tendrá como objetivo **Incrementar el crédito** a través del uso de las tarjetas de crédito que actualmente posee.

Para ello se asumirán los siguientes supuestos:

1. Contar con los datos de contacto del total del grupo objetivo.

El target de la campaña abarca a todos los clientes potencialmente fugados que no tengan uso en sus tarjetas de crédito luego la primera campaña y todos aquellos que además han adquirido una nueva tarjeta de crédito. El enganche de venta para los clientes serán distintos descuentos a través de alianzas estratégicas con restaurantes, tiendas de deportes, tecnología y vestuario.

La alianza con esos comercios será evaluada de acuerdo al uso de la primera campaña, dando énfasis a los convenios más utilizados y agregando nuevas alianzas estratégicas.

Se comunicará a los clientes por distintos canales como SMS, e-mail, llamadas de call center, utilizando la multicanalidad como una oportunidad en la efectividad de la campaña.

TERCER MES: El tercer mes se realizará una gran campaña final que abarca el total de los clientes potencialmente fugados con el objetivo de incrementar el uso de la tarjeta a través del uso de ésta.

- **Campaña 6: Avance en efectivo**

La campaña tendrá como objetivo **Incrementar el crédito** a través del uso de las tarjetas de crédito que actualmente posee.

Para ello se asumirán los siguientes supuestos:

1. Contar con los datos de contacto del total del grupo objetivo.
2. Los clientes aún tienen capacidad de endeudamiento.
3. La evaluación de riesgo de los clientes permite avances en efectivo.

El target de la campaña abarca a todos los clientes potencialmente fugados que tengan o no tengan uso en sus tarjetas de crédito y tengan 1 o más tarjetas, el enganche de venta para los clientes serán tasas preferenciales por avances en efectivo con sus tarjetas, mejores a las tasas de crédito disponibles.

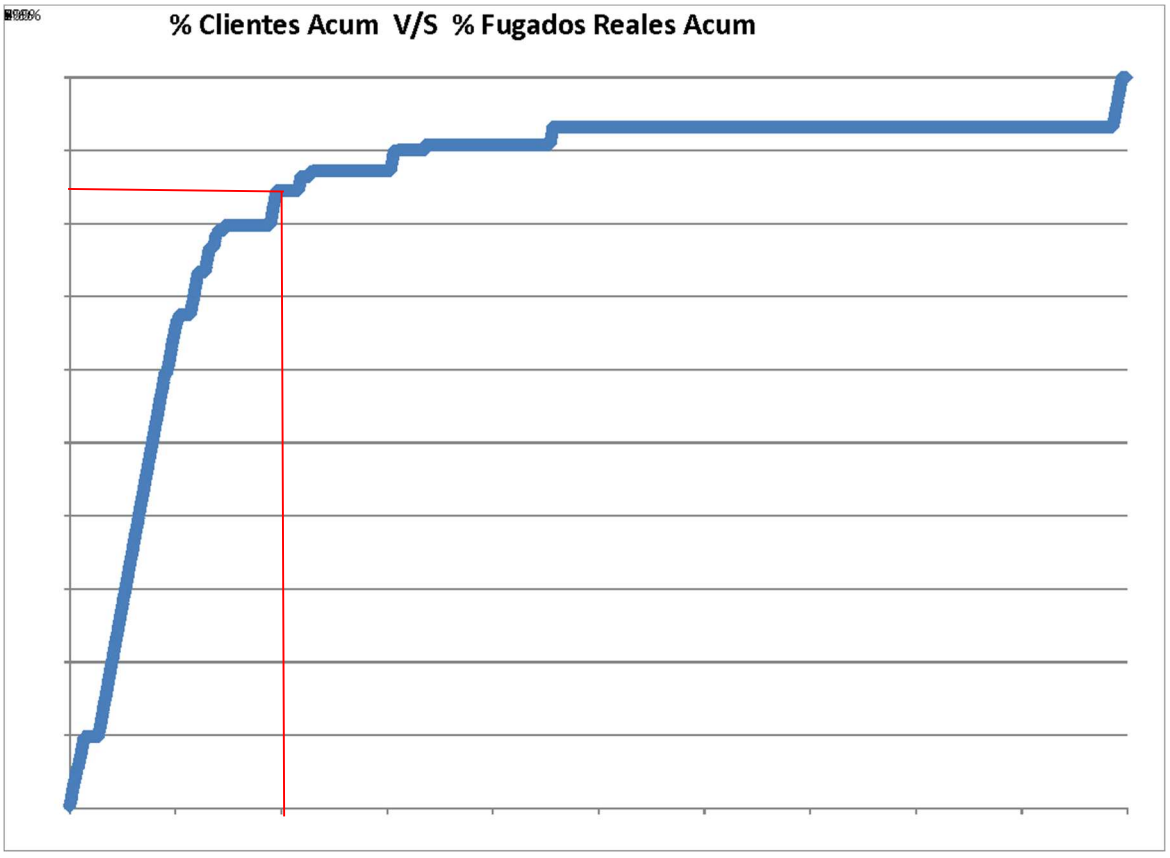
Se comunicará a los clientes por distintos canales como SMS, e-mail, llamadas de call center, utilizando la multicanalidad como una oportunidad en la efectividad de la campaña.

Es relevante considerar que el mercado financiero se encuentra permanentemente enfrentando distintos tipos de campaña impulsadas por la competencia con respecto a igualar y/o mejorar los beneficios que brinda cada institución, además de lidiar con políticas de retención ineficaces, bajas en la calidad de servicios entregados, siendo la cartera de clientes afectada y en consecuencia relacionada directamente con la rentabilidad del negocio.

iii. Evaluación de costos

Todas estas acciones comerciales involucran un conjunto de costos (monetarios, capacidad de empleados, etc.). Como se puede ver en la matriz beneficios/costos descrita en el problema, cualquier acción aplicada a un cliente que previamente es clasificado con alta probabilidad de no fuga, y luego de la acción, decide no fugarse, la acción significa un costo de -100. Para minimizar estos costos, la entidad financiera desea aplicar acciones a los clientes que potencialmente podrían fugarse. Con la clasificación predictiva realizada por el modelo de minería de datos desarrollado, es posible definir el conjunto de clientes sujeto a las acciones comerciales. Por lo tanto, los clientes que sean clasificados por el modelo como posibles fugados, serán el objetivo de las acciones definidas para retenerlos. A continuación se describe de manera detallada la evaluación de costos para las acciones comerciales definidas.

Si se organiza a los clientes desde el más propenso a fugarse de acuerdo al modelo, hasta el menos propenso a fugarse y observamos a cuantos clientes fugados reales abarca, el gráfico sería de la siguiente forma:

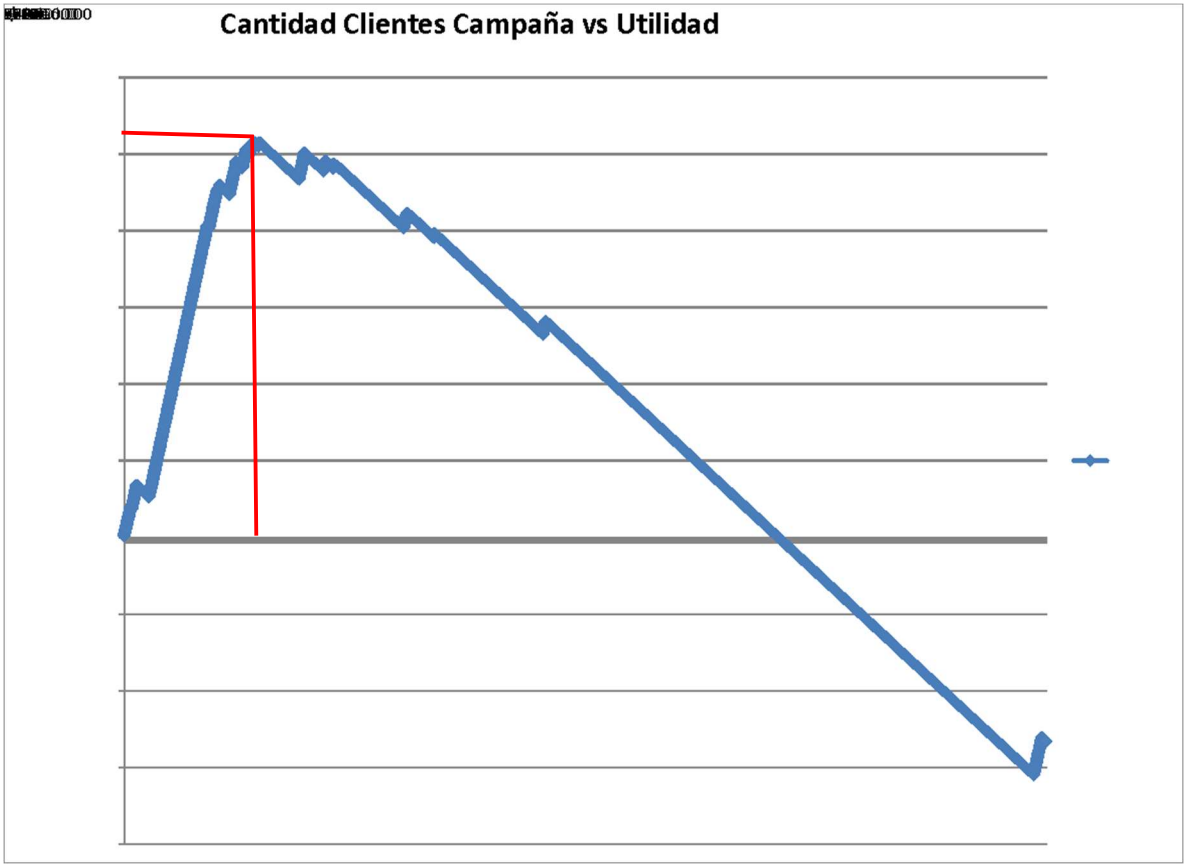


Del gráfico se observa anterior que si se abarca el 20% de los clientes más propensos a fugarse (aprox. 1.100 clientes), se podría identificar cerca de un 85% de los clientes fugados reales (aprox. 600 clientes).

Dado que se busca maximizar las utilidades en una campaña comercial, se debe calcular bajo los valores reales de costos de las campañas, retornos esperados y % de efectividad de las campañas, a continuación, se muestra un caso a modo de ejemplo para explicar el método y así poder identificar el punto máximo de utilidad:

- Si la campaña comercial tiene un costo unitario: $1.000 (Pieza de Marketing)
- Si el retorno del cliente que no se fugó es de: $30.000 (comisión de mantención tarjeta al menos 6 mes posteriores en promedio)
- Si se considera un % Efectividad Campaña de: 20% (por problemas de contactabilidad y/o al cliente no le interesa la campaña)

De esta forma se establece un punto donde se maximicen las utilidades de esta campaña, el gráfico queda de la siguiente forma:



Se observa que existe un punto donde la utilidad de esta campaña de ejemplo se maximiza, este punto se logra con 789 clientes, de los cuales se detectan efectivamente 561 clientes fugados, pero por efecto de la efectividad de la campaña sólo se logran fidelizar a 112 clientes propensos a fugarse (20% efectividad del canal).

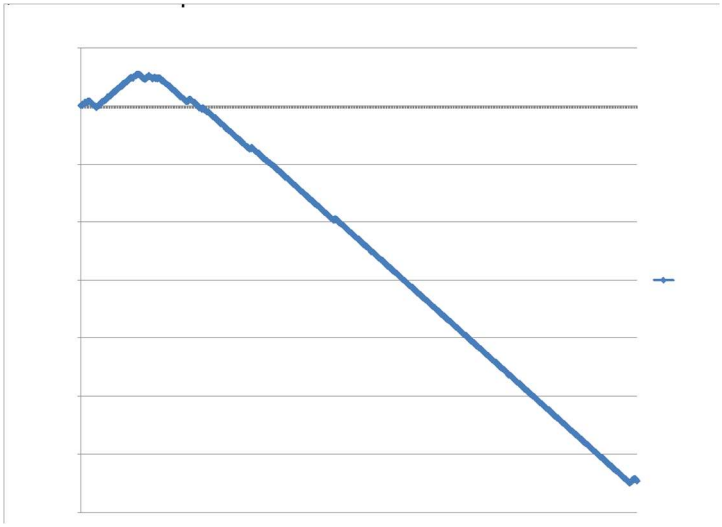
El gasto de la campaña sería de $789.000, y el retorno generado sería de $3.366.000, esto genera una utilidad de $2.577.000, por lo que cualquier otra cantidad de clientes en campaña (superior o inferior a 789 clientes) generaría menos utilidades.

Evaluación de Costos Campañas Reales

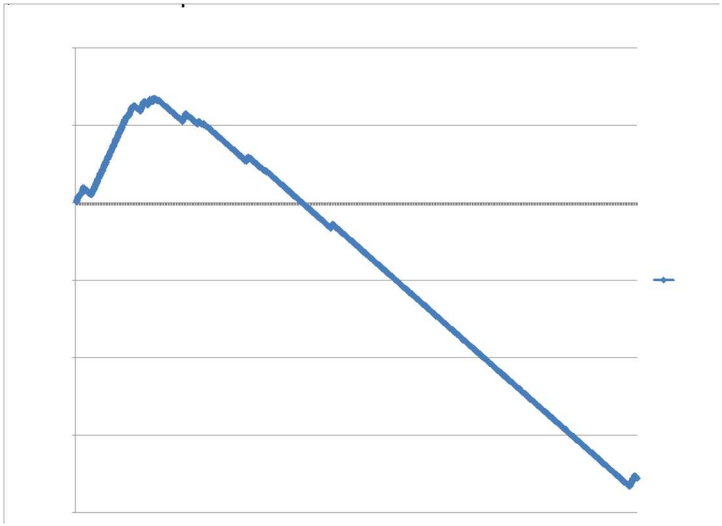
	Gasto Unitario	Retorno Esperado por Cliente Retenido	% Efectividad	Utilidad Campaña Máximizado
Campaña 1	$ 3.072	$ 30.000	20%	$1.086.960
Campaña 2	$ 572	$ 10.000	20%	$ 670.692
Campaña 3	$ 26.101	$ 100.000	35%	$1.072.935
Campaña 4	$ 10.825	$ 60.000	30%	$2.343.875
Campaña 5	$ 572	$ 10.000	20%	$ 670.692

Gráficos de Maximización de Utilidad por Campaña

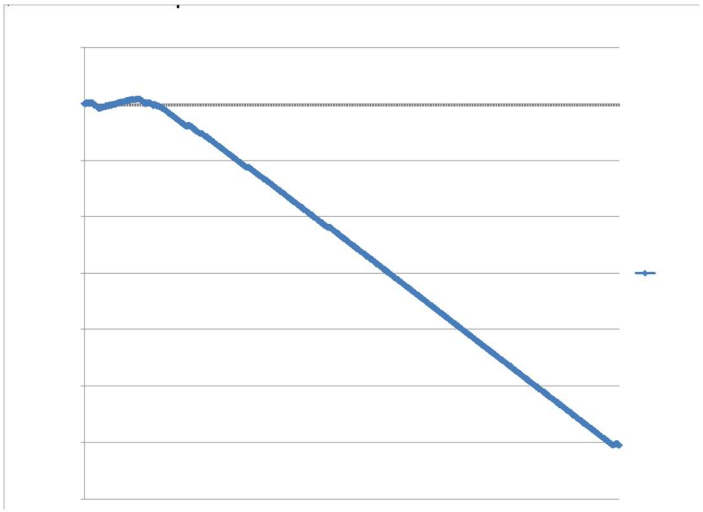
Campaña 1



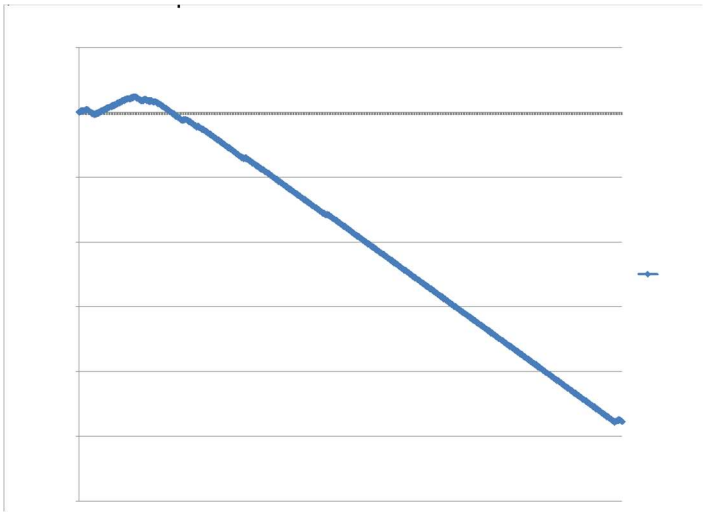
Campaña 2



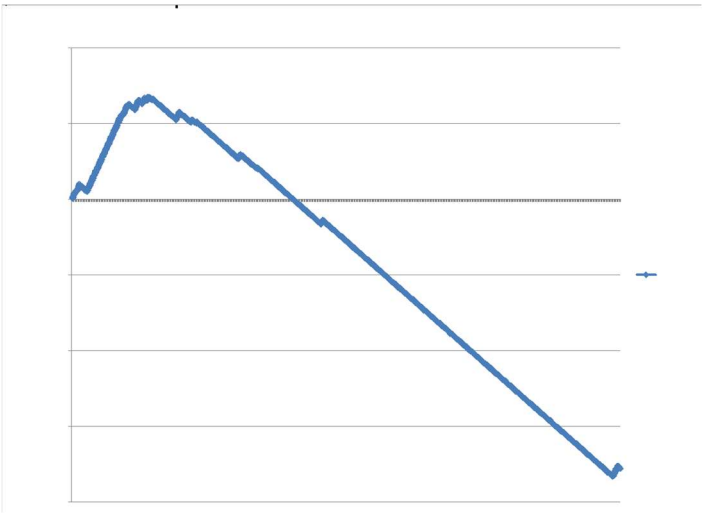
Campaña 3



Campaña 4



Campaña 5



iv. Herramientas tecnológicas

Las herramientas tecnológicas necesarias para desarrollar sistemas de minería de datos que ayudan a resolver los distintos problemas en los procesos de negocio de la entidad financiera son:

- Bases de Datos: Los procesos de una organización deben contar con sistemas de gestión de datos que garanticen su integridad, acceso y procesamiento.
- aplicaciones para empleados y clientes: Algunos de los errores en los datos son causados por el ingreso o edición de éstos en los sistemas usados por empleados y clientes. Es en este contexto en donde se recomienda disponer de aplicaciones para usuarios que incorporen lógicas de validación sobre los datos manipulados.

- herramientas para implementación de modelos predictivos: Para este problema, se utiliza el lenguaje de programación R, el cual permite desarrollar modelos de minería de datos mediante la ejecución de un conjunto de comandos sobre los datos de entrenamiento. Existen distintas tecnologías (gratuitas y pagadas) de análisis y minería de datos a través de operadores y entornos gráficos.

VIII. CONCLUSIONES

La fuga voluntaria de clientes genera significativas pérdidas en cualquier entidad financiera. Para evitarla, periódicamente se definen políticas comerciales enfocadas a la retención, mediante acciones que motiven a los clientes a seguir en la entidad financiera. Para que estas acciones se ejecuten de manera proactiva, es necesario encontrar a los clientes que, por sus características, son posibles fugados. Los modelos predictivos de minería de datos permiten identificar a los clientes propensos a la fuga.

Para los datos incompletos e inconsistentes, se utilizó el reemplazo por media/moda y el reemplazo en base a la relación con otros atributos (en el caso de las variables geográficas). La razón por la cual se realizó este tipo de imputación es el bajo porcentaje de datos erróneos en relación al total, por lo tanto el reemplazo de estos valores no afecta la distribución de los datos. Posterior a este proceso, se realizaron algunas transformaciones sobre los atributos, función logarítmica para las variables monetarias, y variables dummy para variables categóricas, para tener distribuciones sobre los datos del tipo normal, para su posterior normalización (llevar los datos a una escala común).

El uso del test chi-cuadrado, la matriz de correlación y el modelo de regresión logística permiten ordenar los atributos independientes de acuerdo a su relevancia. Esto hizo posible seleccionar los atributos de mayor relevancia para el desarrollo de algunos modelos predictivos de clasificación binaria.

No se observa diferencia significativas en los indicadores comparativos entre los modelos desarrollados. De acuerdo a los indicadores ROC, Sensibilidad y Especificidad, se pueden considerar a todos los modelos como buenos. Hay algunos modelos con alta capacidad predictiva, pero muy complejos de interpretar, como son Red Neuronal y SVM. Ya que se solicitó la interpretación del modelo para la definición de patrones, la selección del modelo predictivo se realizó por el criterio de interpretabilidad, y, según la matriz beneficios/costos, por la Sensibilidad (menor número de falsos negativos). Usando estos criterios, el modelo de Árbol de Decisión fue el seleccionado para ser usado sobre los datos para predicción.

Los patrones fueron definidos en base a las observaciones sobre el modelo predictivo seleccionado. Estos patrones entregan una base, desde el punto de vista del análisis de datos, para la definición de las acciones comerciales.

Las políticas comerciales definidas tienen su base en las características propias del negocio, en la clasificación que realiza el modelo predictivo y en los patrones encontrados. Las tecnologías requeridas apuntan a la implementación adecuada de soluciones de minería de datos, las cuales son un gran apoyo para resolver algunos problemas del negocio de una entidad financiera, como lo es la fuga voluntaria de clientes.

IX. DISCUSIONES

El modelo predictivo realizado para esta tarea usó gran parte de los datos entregados por la entidad financiera, y como se describe en este informe, el modelo fue desarrollado con los atributos que consideró más relevantes . No obstante, otros atributos de los clientes podrían ser considerados para mejorar el modelo y/o para la definición de otros patrones (ejemplo: años con trabajo, años como cliente en la entidad financiera).

En relación con lo anterior, los procesos de negocio y el comportamiento de los clientes cambia con el transcurso del tiempo, y el modelo predictivo a utilizar debe adaptarse a esos cambios. Los datos utilizados para su implementación pueden no ser representativos de la conducta de los clientes más adelante. Es por estos motivos que el modelo debe ser actualizado y ajustado de manera periódica, para mantener (o mejorar) su capacidad predictiva.

Una herramienta utilizada en minería de datos exploratoria es la segmentación de objetos o clustering. Esta técnica permite la agrupación de objetos mediante el análisis comparativo de sus atributos. Si bien para esta tarea se desarrollaron modelos para todos los

clientes, como un futuro desarrollo, se puede considerar una clasificación previa del conjunto de clientes, usando alguna de las metodologías de clustering, para posteriormente desarrollar distintos modelos para cada subconjunto resultante. Con esto, cada segmento de clientes es analizado de manera separada, lo cual permite el desarrollo y uso de modelos que mejor se adapten a las características de cada segmento de clientes.

Las entidades financieras (bancos, cajas de ahorro, cooperativos de crédito, etc.) manejan los beneficios y costos de cada cliente de manera individualizada, de acuerdo a un conjunto de factores (nivel de inversión, créditos asignados, antigüedad, etc.). Para futuras acciones comerciales, es recomendable considerar estos costos y beneficios individualizados, tanto en la clasificación predictiva como en la aplicación de las acciones mismas. Con esto es posible mejorar los beneficios y disminuir costos en la puesta en marcha de políticas comerciales enfocadas a evitar la fuga de clientes.

X. ANEXOS

Anexo 1: Gráficos

Gráfico 1: Diagrama de correlación entre atributos

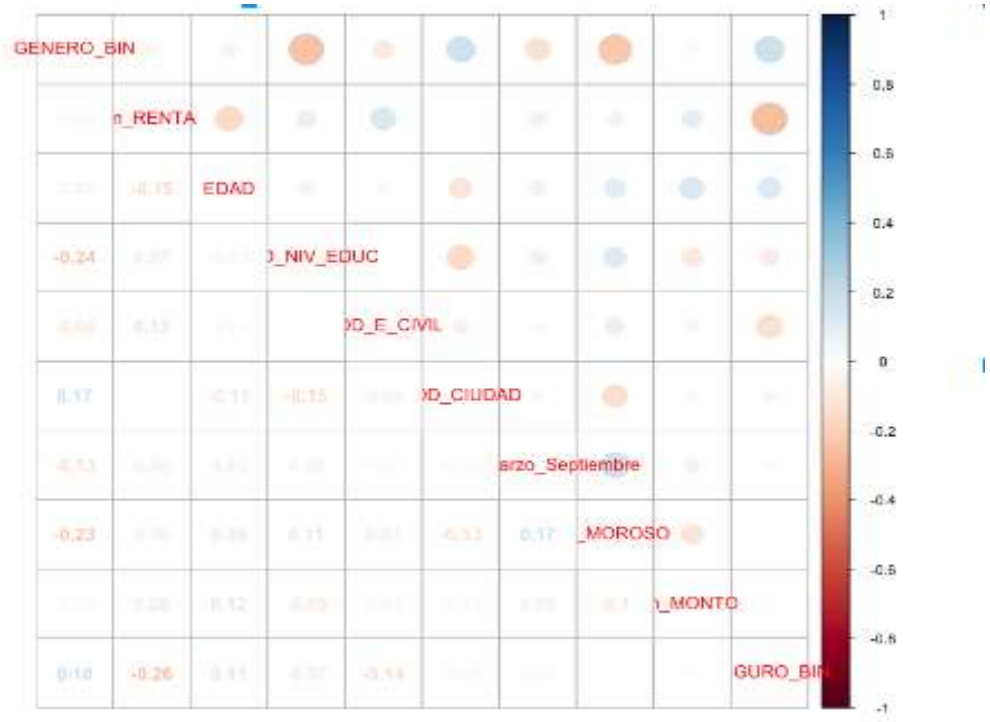


Gráfico 2: K-Nearest Neighbors; K versus ROC

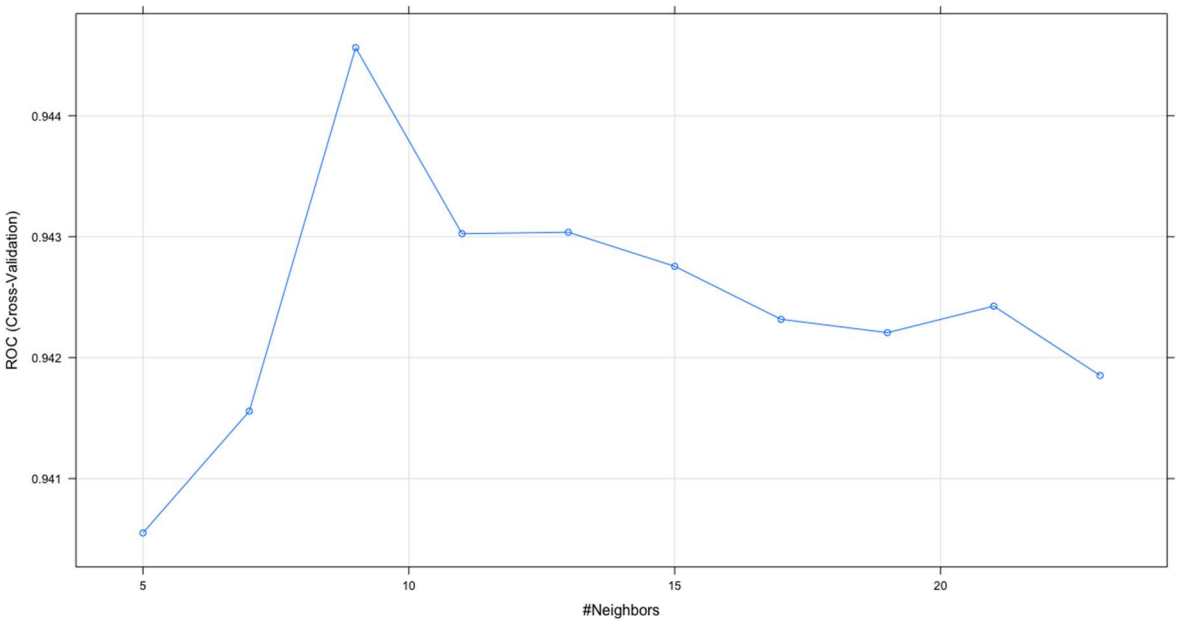


Gráfico 3: Red Neuronal: Nodos en capa escondida y Decaimiento versos ROC

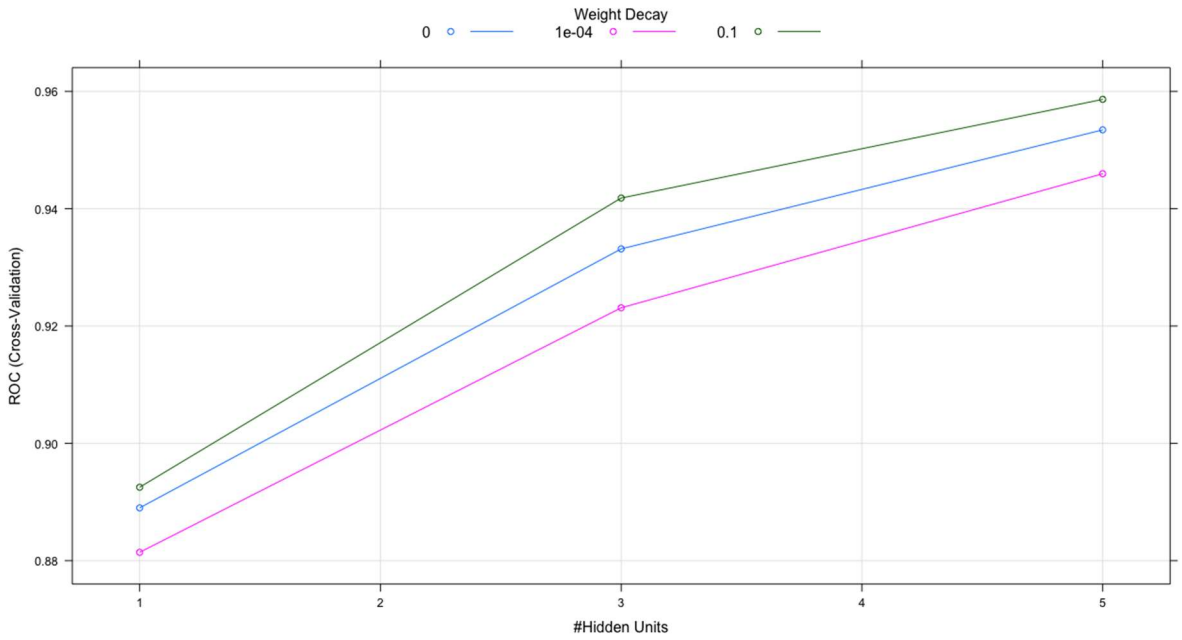
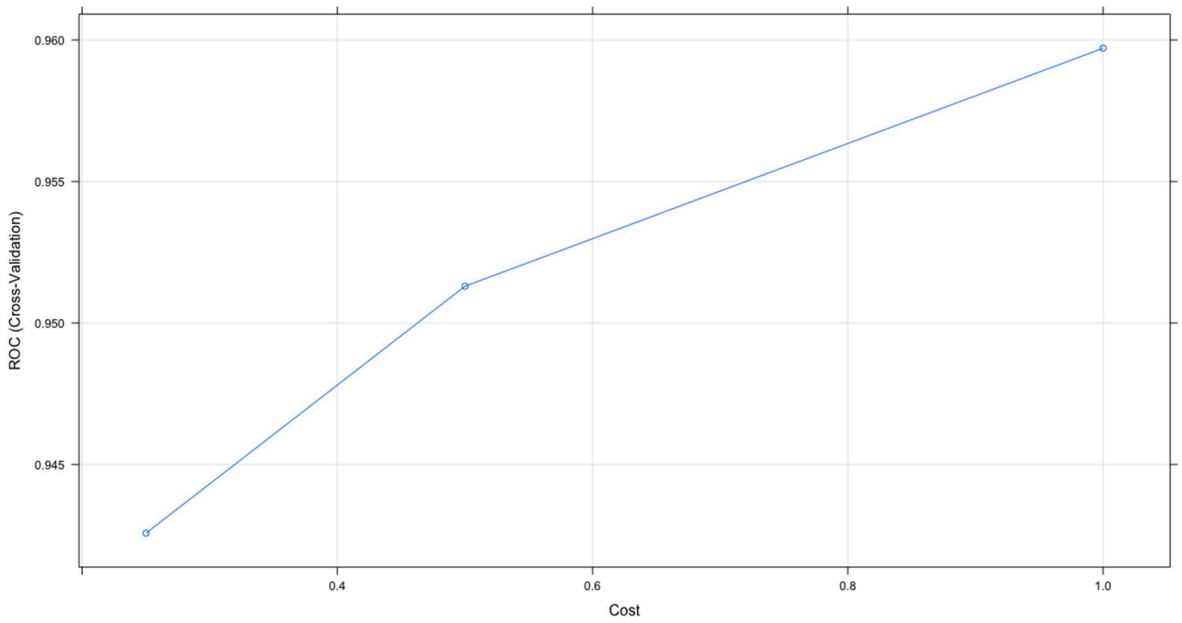


Gráfico 4: Support Vector Machines: Costo versus ROC



Anexo 2: Código R

#tarea diplomado bi

```
#Instalar paquetes (en caso que no esten instalados)
install.packages("readxl")
install.packages("ggplot2")
install.packages("mice")
install.packages("clusterSim")
install.packages("corrplot")
install.packages("MASS")
install.packages("rpart.plot")
install.packages("devtools")
install.packages("WriteXLS")

#Importar paquetes
library(readxl)
library(ggplot2)
library(mice)
library(clusterSim)
library(corrplot)
library(MASS)
library(caret)
library(rpart.plot)
library(devtools)
library(WriteXLS)

#data set
#Cargar BASEFUGA
#Definir carpeta de trabajo y abrir archivo BASEFUGA
#Session -> Set Working Directory -> To source file location
data_set<-read_excel("BASEFUGA.xls", sheet=1)

#Parte 1: AED
summary(data_set)

#Conteo missing values
na_count<-sapply(data_set,function(y)sum(length(which(is.na(y)))))
print(na_count<-data.frame(na_count))

#Parte 2 : Limpieza

#Declarar inconsistencias como valores perdidos
#attach(data_set)
#Edad
data_set$EDAD[data_set$EDAD<0]<-NA
data_set$EDAD[data_set$EDAD>95]<-NA
#Codigo Oficina
data_set$COD_OFI[data_set$COD_OFI==0]<-NA
#NIV_EDUC
data_set$NIV_EDUC[data_set$NIV_EDUC=="EUN"]<-"UNV"



#Imputar Valores Perdidos por media/moda

#metodo para obtener moda
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
```

```
}

#GENERO
data_set$GENERO[is.na(data_set$GENERO)]<-getmode(data_set$GENERO)
#NIV_EDUC
data_set$NIV_EDUC[is.na(data_set$NIV_EDUC)]<-getmode(data_set$NIV_EDUC)
#E_CIVIL
data_set$E_CIVIL[is.na(data_set$E_CIVIL)]<-getmode(data_set$E_CIVIL)
#EDAD
data_set$EDAD[is.na(data_set$EDAD)]<-mean(data_set$EDAD,na.rm=TRUE)
#COD_OFI
data_set$COD_OFI[is.na(data_set$COD_OFI)]<-
getmode(data_set$COD_OFI[data_set$CIUDAD=='VALPARAISO' & data_set$COD_COM==66])
#COD_COM
data_set$COD_COM[is.na(data_set$COD_COM) & data_set$COD_OFI==45]<-
getmode(data_set$COD_COM[data_set$COD_OFI==45])
data_set$COD_COM[is.na(data_set$COD_COM) & data_set$COD_OFI==90]<-
getmode(data_set$COD_COM[data_set$COD_OFI==90])
#CIUDAD
data_set$CIUDAD[is.na(data_set$CIUDAD) & data_set$COD_OFI==45 & data_set$COD_COM==88]<-
getmode(data_set$CIUDAD[data_set$COD_OFI==45 & data_set$COD_COM==88])
data_set$CIUDAD[is.na(data_set$CIUDAD) & data_set$COD_OFI==90 & data_set$COD_COM==80]<-
getmode(data_set$CIUDAD[data_set$COD_OFI==90 & data_set$COD_COM==80])

#Graficos
#ggplot(data_set, aes(x=GENERO, fill=FUGA))+geom_bar()
#ggplot(data_set, aes(x=NIV_EDUC, fill=FUGA))+geom_bar()
#ggplot(data_set, aes(x=E_CIVIL, fill=FUGA))+geom_bar()
#ggplot(data_set, aes(x=M_MOROSO, fill=FUGA))+geom_bar()
#ggplot(data_set, aes(x=SEGURO, fill=FUGA))+geom_bar()
#ggplot(data_set, aes(x=RENTA))+geom_histogram(na.rm=TRUE)
#ggplot(data_set, aes(x=D_Marzo))+geom_histogram(na.rm=TRUE)
#ggplot(data_set, aes(x=D_Abril))+geom_histogram(na.rm=TRUE)
#ggplot(data_set, aes(x=D_Mayo))+geom_histogram(na.rm=TRUE)
#ggplot(data_set, aes(x=D_Junio))+geom_histogram(na.rm=TRUE)
#ggplot(data_set, aes(x=D_Julio))+geom_histogram(na.rm=TRUE)
#ggplot(data_set, aes(x=D_Agosto))+geom_histogram(na.rm=TRUE)
#ggplot(data_set, aes(x=D_Septiembre))+geom_histogram(na.rm=TRUE)
#ggplot(data_set, aes(x=MONTO))+geom_histogram(na.rm=TRUE)

#Parte 3: Transformacion
attach(data_set)

#FUGA
FUGA<-ifelse(FUGA=="FUGA","FUGA","NO_FUGA")




#Variables 'Dummy'
#GENERO_BIN: 1->F, 0->M
GENERO_BIN<-as.numeric(GENERO=="F")
#SEGURO_BIN: 1->SI, 0->NO
SEGURO_BIN<-as.numeric(SEGURO=="SI")

#RENTA
```

```
ln_RENTA<-log(RENTA + 1)
#ggplot(data_set, aes(x=ln_RENTA))+geom_histogram(na.rm=TRUE)

#D_Marzo hasta D_Septiembre
D_Marzo_Septiembre<-(D_Marzo+D_Abril+D_Mayo+D_Junio+D_Julio+D_Agosto+D_Septiembre)
ln_D_Marzo_Septiembre<-log(D_Marzo_Septiembre+1)
#ggplot(data_set, aes(x=ln_D_Marzo_Septiembre))+geom_histogram(na.rm=TRUE)

#MONTO
ln_MONTO<-log(MONTO+1)
#ggplot(data_set, aes(x=ln_MONTO))+geom_histogram(na.rm=TRUE)

#CIUDAD
#COD_CIUDAD: 0->SANTIAGO, 1->ARICA, 2->CONCEPCION, 3->Otras
COD_CIUDAD<-
ifelse(CIUDAD=="SANTIAGO",0,ifelse(CIUDAD=="ARICA",1,ifelse(CIUDAD=="CONCEPCION",2,3)))
#ggplot(data_set, aes(x=COD_CIUDAD, fill=FUGA))+geom_bar()

#NIV_EDUC
#COD_NIV_EDUC: 0->"BAS" o "MED", 1->"TEC", 2->"UNV"
COD_NIV_EDUC<-ifelse((NIV_EDUC=="BAS" | NIV_EDUC=="MED"),0,ifelse(NIV_EDUC=="TEC",1,2) )
#ggplot(data_set, aes(x=COD_NIV_EDUC, fill=FUGA))+geom_bar()

#E_CIVIL
#COD_E_CIVIL: 0->"SOL", 1->"CAS", 2->"VIU", 3->"SEP"
COD_E_CIVIL<-ifelse(E_CIVIL=="SOL",0,ifelse(E_CIVIL=="CAS",1,ifelse(E_CIVIL=="VIU",2,3)))
#ggplot(data_set, aes(x=COD_E_CIVIL, fill=FUGA))+geom_bar()

#Normalizaci?n: Escalamiento 0-1
data_set_T<-data.frame(GENERO_BIN, ln_RENTA, EDAD, COD_NIV_EDUC, COD_E_CIVIL, COD_CIUDAD,
                       ln_D_Marzo_Septiembre,M_MOROSO, ln_MONTO,SEGURO_BIN)

data_set_T<-data.Normalization(data_set_T,type="n4",normalization="column")

sapply(data_set_T, sd)

#Redundancia

CorrMat<-cor(data_set_T)
corrplot.mixed(CorrMat)

#Relevancia
#GENERO
tbl<-table(GENERO_BIN, FUGA)
tbl
chisq.test(tbl)



#EDAD es num?rica -> se debe discretizar primero
bins<-5
cutpoints<-quantile(EDAD,(0:bins)/bins)
EDAD_binned <-cut(EDAD,cutpoints,include.lowest=TRUE)
tbl<-table(EDAD_binned, FUGA)
chisq.test(tbl)

#NIV_EDUC
```

```
tbl<-table(COD_NIV_EDUC, FUGA)
chisq.test(tbl)

#E_CIVIL
tbl<-table(COD_E_CIVIL, FUGA)
chisq.test(tbl)

#CIUDAD
tbl<-table(COD_CIUDAD, FUGA)
chisq.test(tbl)

#M_MOROSO
tbl<-table(M_MOROSO, FUGA)
chisq.test(tbl)

#SEGURO
tbl<-table(SEGURO, FUGA)
chisq.test(tbl)

#RENTA es num?rica -> se debe discretizar primero
bins<-5
cutpoints<-quantile(RENTA,(0:bins)/bins)
RENTA_binned <-cut(RENTA,cutpoints,include.lowest=TRUE)
tbl<-table(RENTA_binned, FUGA)
chisq.test(tbl)

#D_Marzo_Septiembre es num?rica -> se debe discretizar primero
bins<-5
cutpoints<-quantile(D_Marzo_Septiembre,(0:bins)/bins)
D_Marzo_Septiembre_binned <-cut(D_Marzo_Septiembre,cutpoints,include.lowest=TRUE)
tbl<-table(D_Marzo_Septiembre_binned, FUGA)
chisq.test(tbl)

#MONTO es num?rica -> se debe discretizar primero
bins<-5
cutpoints<-quantile(MONTO,(0:bins)/bins)
MONTO_binned <-cut(MONTO,cutpoints,include.lowest=TRUE)
tbl<-table(MONTO_binned, FUGA)
chisq.test(tbl)


#Descartar SEGURO y RENTA
data_set_T<-subset(data_set_T, select = -c(SEGURO_BIN,ln_RENTA))
data_set_Model<-data.frame(data_set_T,FUGA)




#Regresion Logistica para seleccion de atributos
logreg_fit<-train(FUGA ~ ., data = data_set_Model, method="glm", family="binomial")
summary(logreg_fit)

data_set_Val<-data_set_Model

#Parte 5: VALIDACION
# Validaci??n y comparaci??n de modelos
#Cross-Validation
```

```
ctrl<-trainControl(method="cv",number= 10, summaryFunction=twoClassSummary, classProbs = TRUE)
#ctrl<-trainControl(method="cv",number= 10, summaryFunction=twoClassSummary, sampling="smote",
classProbs = TRUE)
#ctrl<-trainControl(method="cv",number= 10, summaryFunction=twoClassSummary, sampling="down",
classProbs = TRUE)
#Regresion Logistica
set.seed(825)
logreg_fit<-train(FUGA ~ ., data = data_set_Val, method="glm", family="binomial", trControl = ctrl,
metric="ROC")
logreg_fit

#k-NN
set.seed(825)
knn_fit<-train(FUGA ~ ., data = data_set_Val, method = "knn",tuneLength = 10, trControl = ctrl, metric="ROC")
knn_fit
plot(knn_fit)

# Arbol no optimizado
set.seed(825)
tree_fit<-train(FUGA ~ ., data = data_set_Val, method = "rpart", trControl = ctrl,
metric="ROC",tuneGrid=data.frame(cp=0.009))
tree_fit
tree_fit$finalModel
rpart.plot(tree_fit$finalModel)

#Neural Net
set.seed(825)
nnet_fit<-train(FUGA ~ ., data = data_set_Val, method = "nnet", trControl = ctrl, metric="ROC")
nnet_fit
plot(nnet_fit)
#Nota: el siguiente codigo requiere internet
source_url('https://gist.githubusercontent.com/fawda123/7471137/raw/466c1474d0a505ff044412703516c34
f1a4684a5/nnet_plot_update.r')
plot.nnet(nnet_fit,nid=T)

#SVM (kernel)
set.seed(825)
SVM_fit<-train(FUGA ~ ., data = data_set_Val, method = "svmRadial", trControl = ctrl, metric="ROC")
SVM_fit
plot(SVM_fit)


#Graficar Resultados
results <- resamples(list(LOGIT=logreg_fit, KNN=knn_fit, TREE=tree_fit, NN=nnet_fit, SVM=SVM_fit))
summary(results)
# boxplots of results
bwplot(results)
##################################################
#Predicci??n
data_set_pred<-read_excel("BASEFUGA.xls", sheet=2)
summary(data_set_pred)

#Conteo missing values
na_count<-sapply(data_set_pred,function(y)sum(length(which(is.na(y)))))
print(na_count<-data.frame(na_count))
```

```
data_set_pred$EDAD[data_set_pred$EDAD>95]<-NA

#Imputar Valores Perdidos por media/moda

#metodo para obtener moda
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

#GENERO
data_set_pred$GENERO[is.na(data_set_pred$GENERO)]<-getmode(data_set_pred$GENERO)
#NIV_EDUC
data_set_pred$NIV_EDUC[is.na(data_set_pred$NIV_EDUC)]<-getmode(data_set_pred$NIV_EDUC)
#E_CIVIL
data_set_pred$E_CIVIL[is.na(data_set_pred$E_CIVIL)]<-getmode(data_set_pred$E_CIVIL)
#EDAD
data_set_pred$EDAD[is.na(data_set_pred$EDAD)]<-mean(data_set_pred$EDAD,na.rm=TRUE)
#M_MOROSO
data_set_pred$M_MOROSO[is.na(data_set_pred$M_MOROSO)]<-
mean(data_set_pred$M_MOROSO,na.rm=TRUE)
#COD_COM
data_set_pred$COD_COM[is.na(data_set_pred$COD_COM) & data_set_pred$COD_OFI==90]<-
getmode(data_set_pred$COD_COM[data_set_pred$COD_OFI==90])
#COD_CIUDAD
data_set_pred$CIUDAD[is.na(data_set_pred$CIUDAD) & data_set_pred$COD_OFI==90 &
data_set_pred$COD_COM==89]<-getmode(data_set_pred$CIUDAD[data_set_pred$COD_OFI==90 &
data_set_pred$COD_COM==89])
data_set_pred$CIUDAD[is.na(data_set_pred$CIUDAD) & data_set_pred$COD_OFI==56 &
data_set_pred$COD_COM==131]<-getmode(data_set_pred$CIUDAD[data_set_pred$COD_OFI==56 &
data_set_pred$COD_COM==131])
data_set_pred$CIUDAD[is.na(data_set_pred$CIUDAD) & data_set_pred$COD_OFI==27 &
data_set_pred$COD_COM==226]<-getmode(data_set_pred$CIUDAD[data_set_pred$COD_OFI==27 &
data_set_pred$COD_COM==226])
data_set_pred$CIUDAD[is.na(data_set_pred$CIUDAD) & data_set_pred$COD_OFI==55 &
data_set_pred$COD_COM==90]<-getmode(data_set_pred$CIUDAD[data_set_pred$COD_OFI==55 &
data_set_pred$COD_COM==90])
data_set_pred$CIUDAD[is.na(data_set_pred$CIUDAD) & data_set_pred$COD_OFI==138 &
data_set_pred$COD_COM==1]<-getmode(data_set_pred$CIUDAD[data_set_pred$COD_OFI==138 &
data_set_pred$COD_COM==1])
data_set_pred$CIUDAD[is.na(data_set_pred$CIUDAD) & data_set_pred$COD_COM==253]<-
getmode(data_set_pred$CIUDAD[data_set_pred$COD_COM==253])
data_set_pred$CIUDAD[is.na(data_set_pred$CIUDAD) & data_set_pred$COD_OFI==43 &
data_set_pred$COD_COM==70]<-getmode(data_set_pred$CIUDAD[data_set_pred$COD_OFI==43 &
data_set_pred$COD_COM==70])


#Transformacion
attach(data_set_pred)
GENERO_BIN<-as.numeric(GENERO=="F")
D_Marzo_Septiembre<-(D_Marzo+D_Abril+D_Mayo+D_Junio+D_Julio+D_Agosto+D_Septiembre)
ln_D_Marzo_Septiembre<-log(D_Marzo_Septiembre+1)
ln_MONTO<-log(MONTO+1)
COD_CIUDAD<-
ifelse(CIUDAD=="SANTIAGO",0,ifelse(CIUDAD=="ARICA",1,ifelse(CIUDAD=="CONCEPCION",2,3)))
COD_NIV_EDUC<-ifelse((NIV_EDUC=="BAS" | NIV_EDUC=="MED"),0,ifelse(NIV_EDUC=="TEC",1,2) )
COD_E_CIVIL<-ifelse(E_CIVIL=="SOL",0,ifelse(E_CIVIL=="CAS",1,ifelse(E_CIVIL=="VIU",2,3)))
```

```
data_set_pred_T<-data.frame(GENERO_BIN, EDAD, COD_NIV_EDUC,
COD_E_CIVIL,COD_CIUDAD,ln_D_Marzo_Septiembre, M_MOROSO, ln_MONTO)
data_set_pred_T<-data.Normalization(data_set_pred_T,type="n4",normalization="column")

#ARBOL DE DECISION para prediccion
y_pred<-predict(tree_fit, newdata=data_set_pred_T)
View(data.frame(y_pred))
summary(y_pred)

#mostrar probabilidad
y_pred_prob<-predict(tree_fit, newdata=data_set_pred_T,type="prob")
View(data.frame(y_pred_prob))

#agregar columna de prediccion a validacion
data_set_pred_final<-data.frame(data_set_pred)
data_set_pred_final$FUGA<-y_pred
colnames(data_set_pred_final)[20]<-"FUGA"

#Pegar datos de validacion con prediccion a Base_clasificados.xls
WriteXLS(data_set_pred_final,"Base_clasificados.xls")
```