

Data Mining Practical Exercise (Homework)



Minería de datos

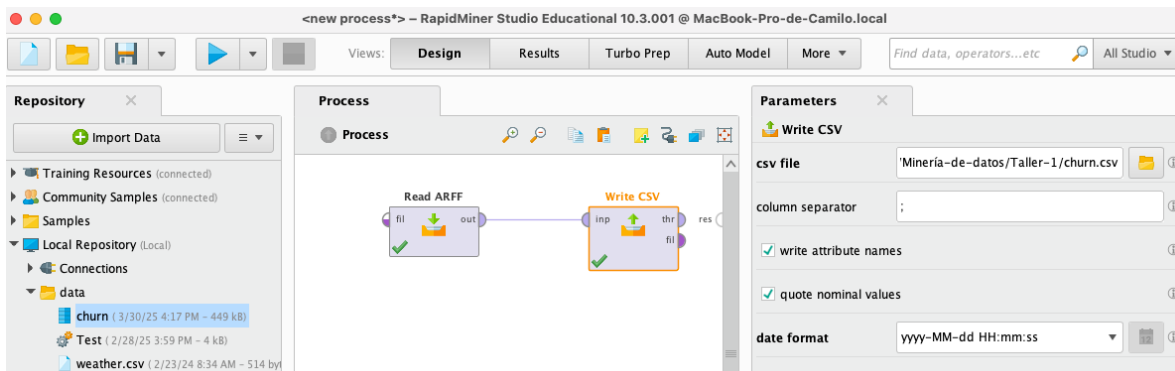
Presentado por: Camilo Enrique Romero Parra

Presentado a: PhD. Carlos Alberto Cobos Lozada

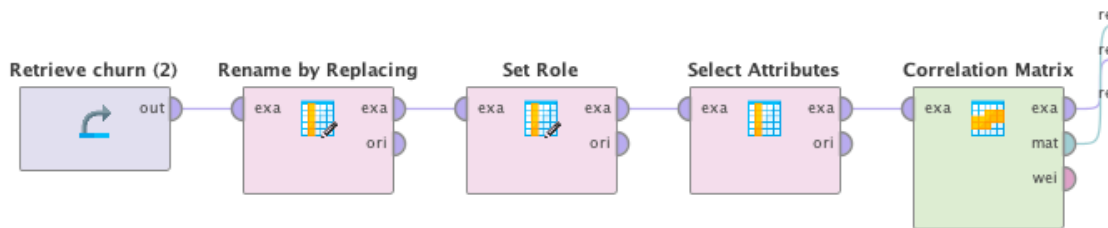
Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Programa de Doctorado en Ciencias de la Computación
Popayán, Cauca
2025

1. Análisis exploratorio de datos

Utilizando RapidMiner se realiza inicialmente la lectura de los datos con el archivo .arff.



Con los datos en formato csv, se realiza una preparación de los datos, cambiando el nombre de las columnas para eliminar espacios en blanco y definir el rol de la columna “churn” como label y la columna “Phone Number” como id. Adicionalmente, se eliminan los atributos de “State” y “Area Code”, dado que dentro del conocimiento del contexto se sabe que estos atributos tienen información errónea.



Adicionalmente, se mapearon los valores de Inter plan y de voice mail plan, dejando los valores de “no” como 0 y “yes” como 1.

1.1. Variables posiblemente correlacionadas

Utilizando la matriz de correlación, se obtuvo la visualización del mapa de calor:



Es posible observar la correlación entre las variables que se muestran en la tabla a continuación:

Variable 1	Variable 2	Coeficiente de correlación
No_of_Vmail_Mesgs	VoiceMail_Plan	0.957
Total_Day_Min	Total_Day_Charge	1
Total_Evening_Min	Total_Evening_Charge	1
Total_Night_Minutes	Total_Night_Charge	1
Total_Int_Min	Total_Int_Charge	1

Con base en la matriz de correlación podemos afirmar que los minutos que consume el usuario, ya sea de día, tarde, o noche, o de manera internacional, tienen una correlación lineal perfecta con respecto al costo que la compañía telefónica carga al usuario.

Adicionalmente, se puede observar una correlación lineal fuerte entre la variable “VoiceMail Plan” que indica si el usuario tiene o no un plan de mensajes de voz y la variable “No of Vmail Mesgs” que indica el número de mensajes de voz. Por lo tanto, se puede inferir que la presencia de un plan de voz esta directamente relacionada con el número de mensajes de voz que el usuario envíe.

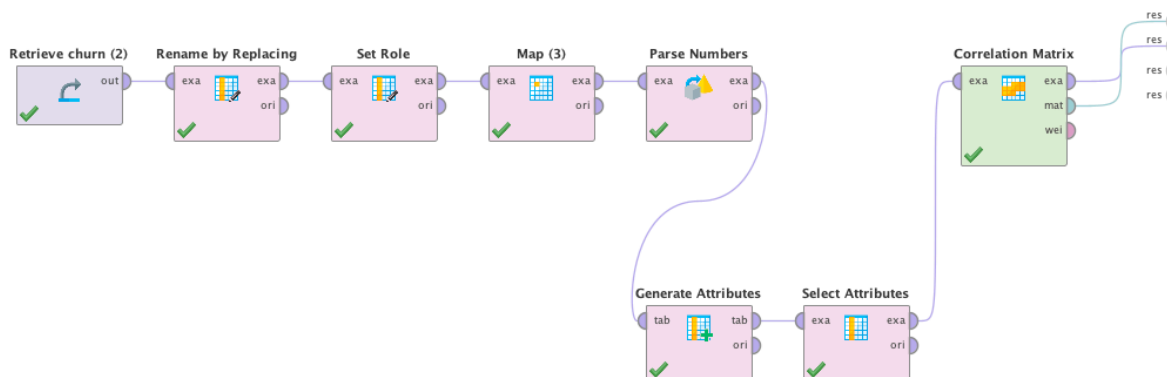
1.2. Variables eliminadas

Con base en el análisis realizado se puede tomar la decisión de eliminar una de las dos variables que estén correlacionadas linealmente de manera perfecta. Esto simplifica el modelo y puede mejorar su rendimiento y reducir el riesgo de sobreajuste. Por este motivo, se eliminarán las siguientes variables: Total_Day_Charge, Total_Evening_Charge, Total_Night_Charge y Total_Int_Charge. Sin embargo, es necesario mantener el valor de cargo total del usuario, por lo que se añade la variable Total_Charge que será la suma de los otros cargos.

Con respecto al atributo No_of_Vmail_Mesgs, este será cero siempre que el atributo "VoiceMail Plan" sea "no". Es posible crear un nuevo atributo que una la información de ambas variables (N_of_Vmail_Mesgs_VMail_Plan). Para hacer esto es necesario mapear VoiceMail_Plan para que pase de ser una variable binomial a una variable numérica. Una vez ambas variables son numéricas se puede aplicar la siguiente condicional:

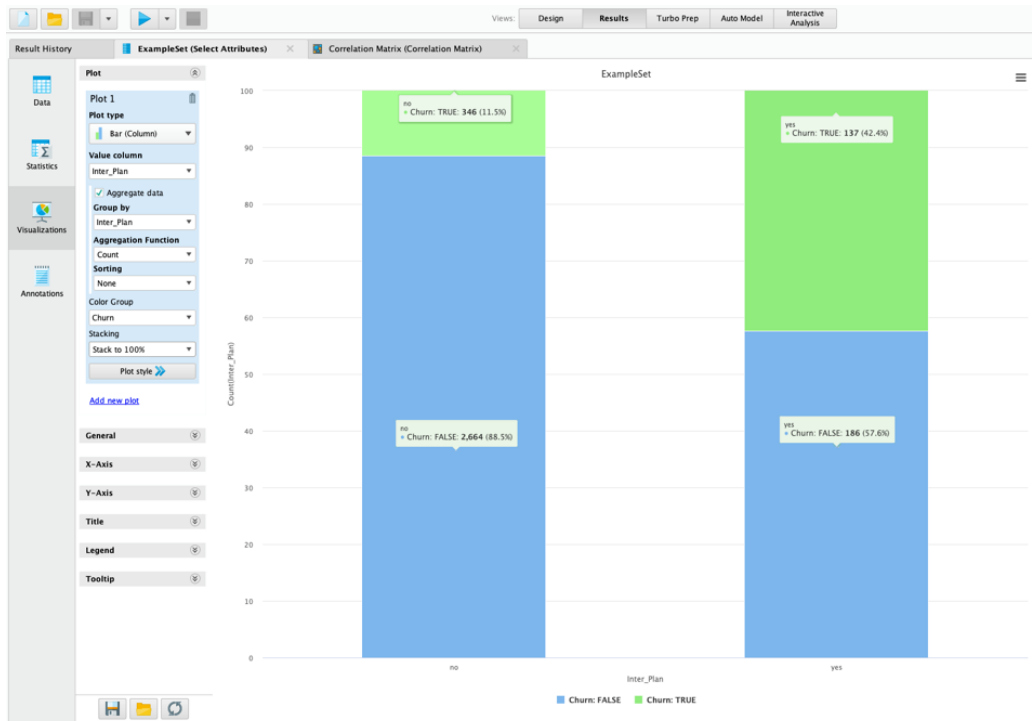
$$\text{if}(\text{VoiceMail_Plan}==0,-1,\text{No_of_Vmail_Mesgs})$$

De tal manera que, si el usuario no tiene plan de mensajes de voz, el número de mensajes que envíe sea -1. Ahora bien, si el número de mensajes es de 0, significa que la persona sí tenía plan de mensajes de voz pero no lo utilizaba. De esta manera el atributo VoiceMail_Plan puede ser eliminado.



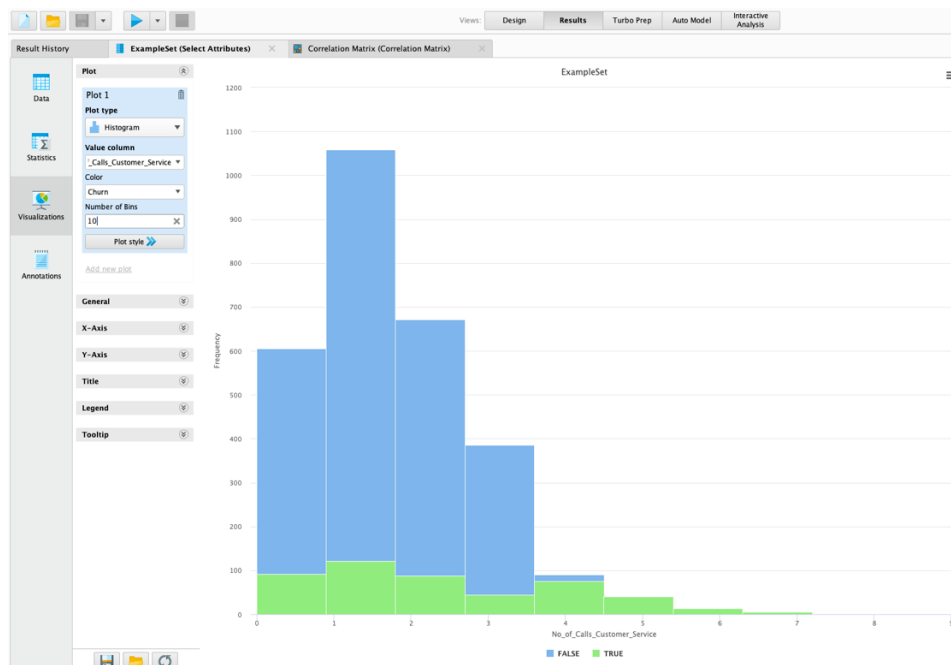
1.3. Churners vs Inter Plan

Al analizar el impacto de la variable Inter_Plan con la variable objetivo Churn, se observa que el 42,4% de los usuarios con plan internacional decidieron cambiar de empresa de telefonía, esto implica que el plan tiene un gran impacto en la decisión final del usuario.

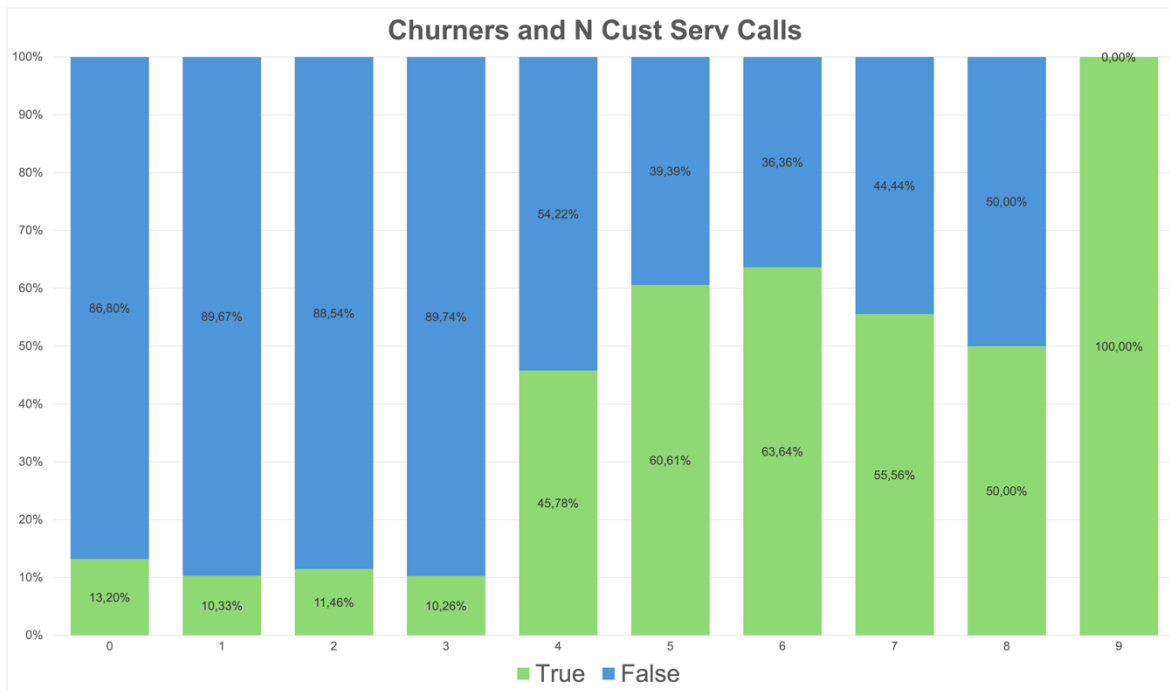


1.4. Churners vs number of calls to customer service

Al observar el grafico que relaciona el número de llamadas que realiza un usuario al servicio técnico con la variable objetivo Churn, se aprecia que, después de la tercera llamada a servicio técnico los usuarios tienen una tendencia muy alta a cambiar de compañía telefónica. Se puede concluir que un número de llamadas a servicio técnico superior a 3 tiene una gran influencia en la decisión del cliente.

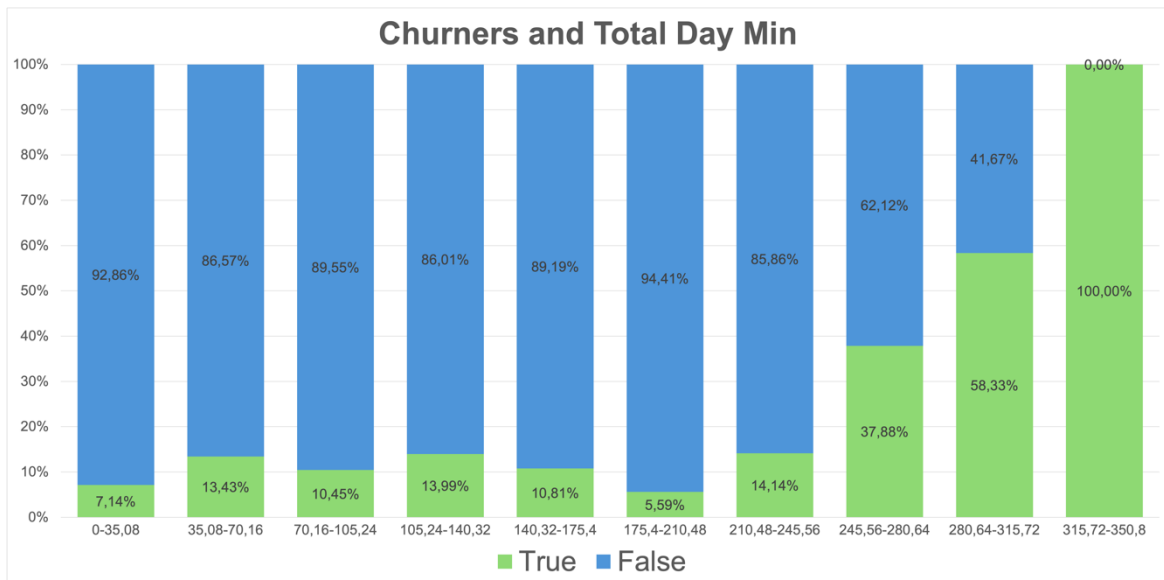


Dado que RapidMiner no permite realizar histogramas que muestren la frecuencia por categorías de una variable con dos colores distintos (en este caso para señalar el Churn) y que a su vez permita hacer stack de manera porcentual, es necesario realizar el grafico con otra herramienta estadística, en este caso Excel. En este grafico se puede apreciar más fácilmente que, después de la cuarta llamada que se realiza a servicio técnico, el número de usuarios que cambian de compañía aumenta hasta 45% en la cuarta llamada, llegando a ser del 100% en la novena llamada.



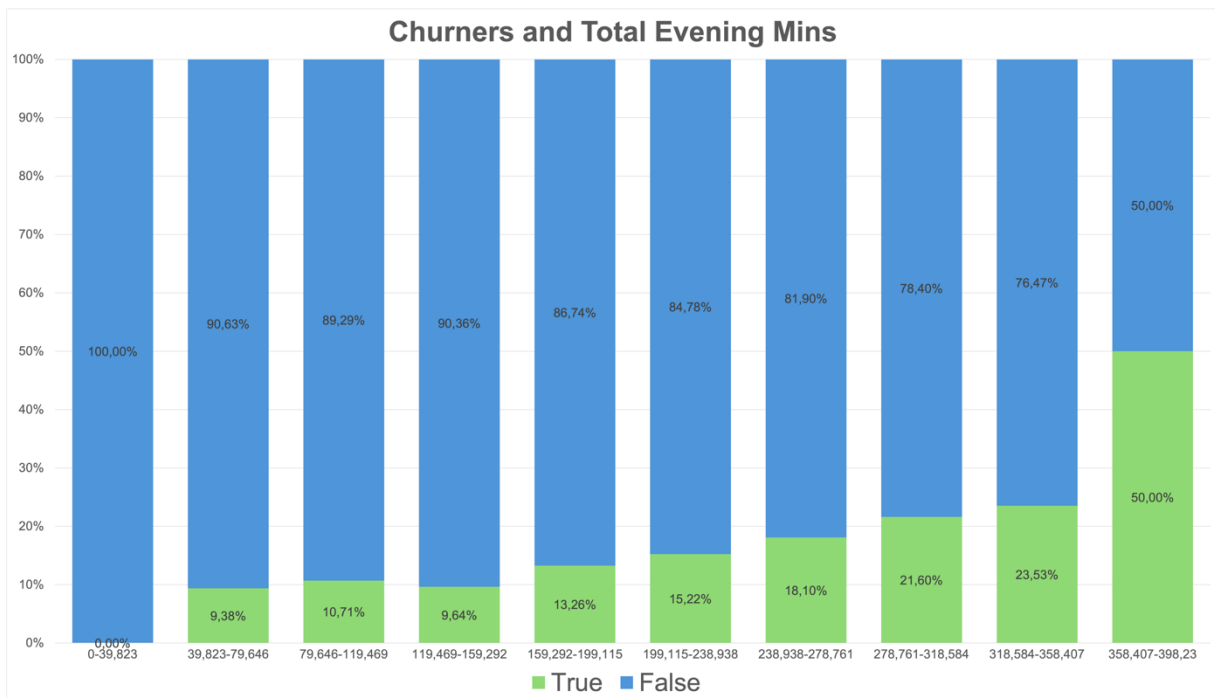
1.5. Churners vs Total Day Min

Al observar el grafico que relaciona los minutos de llamadas que realiza un usuario al día con la variable objetivo Churn, se puede concluir que, al realizar más de 245 minutos de llamadas al día el 37,88% de los clientes cambian de compañía telefónica. Además de esto, por encima de 315 llamadas todos los clientes cambiaron de compañía.



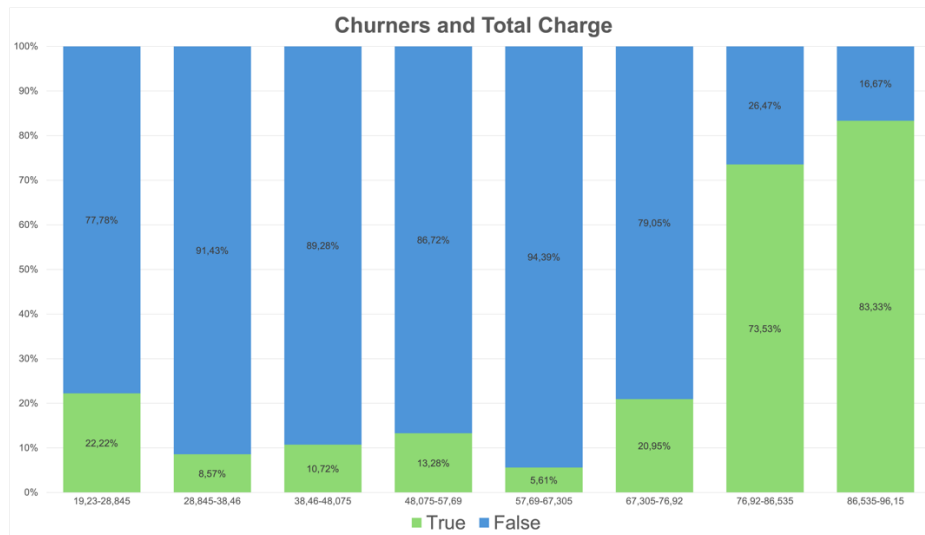
1.6. Churners vs Total Evening Min

Al observar el grafico que relaciona los minutos de llamadas que realiza un usuario en la tarde-noche con la variable objetivo Churn, con curva de campana, se puede concluir que, al realizar más de 358 minutos de llamadas en la tarde-noche la mitad (50%) de los clientes cambian de compañía telefónica.

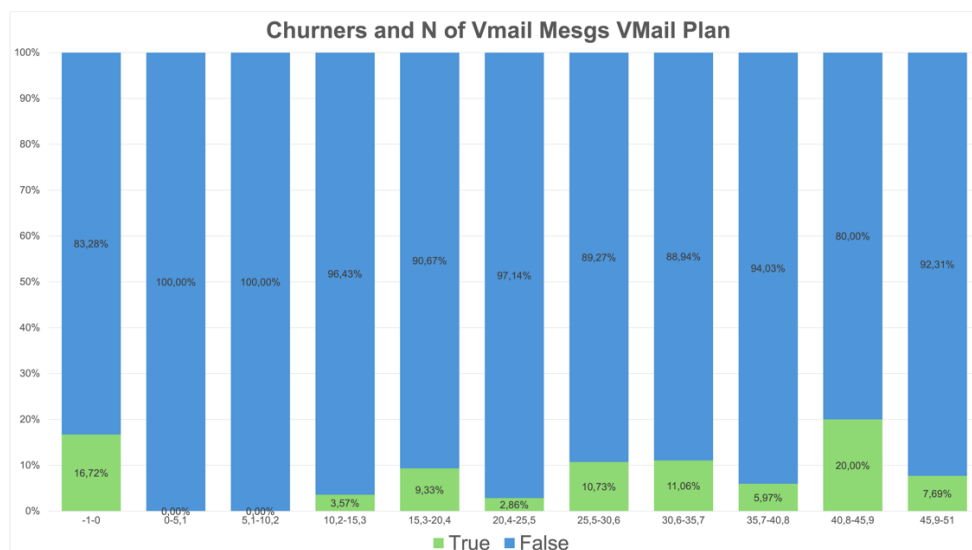


1.7. Análisis del resto de variables vs Churn

Con respecto a los minutos totales consumidos en las noches y en llamadas internacionales, no se encuentra un valor que destaque en comparación a las relaciones antes vistas. Por otro lado, al observar el atributo Total Charge es posible observar que, cuando el cargo supera los 76,92, se aprecia un incremento significativo en el número de usuarios que se retiran del servicio telefónico (de 73,53%), como se muestra en el grafico a continuación:



Con respecto a la variable que creada para almacenar la información del número de mensajes de voz y los usuarios con plan de voz, es posible apreciar que aquellos usuarios sin plan de voz (categoría -1-0) se han retirado de la compañía en un 16,72%. Sin embargo, las personas con plan de mensajes de voz tiende a permanecer en el servicio, presentando un pico de migración entre 40 y 46 mensajes con un 20% de retiro de la compañía:



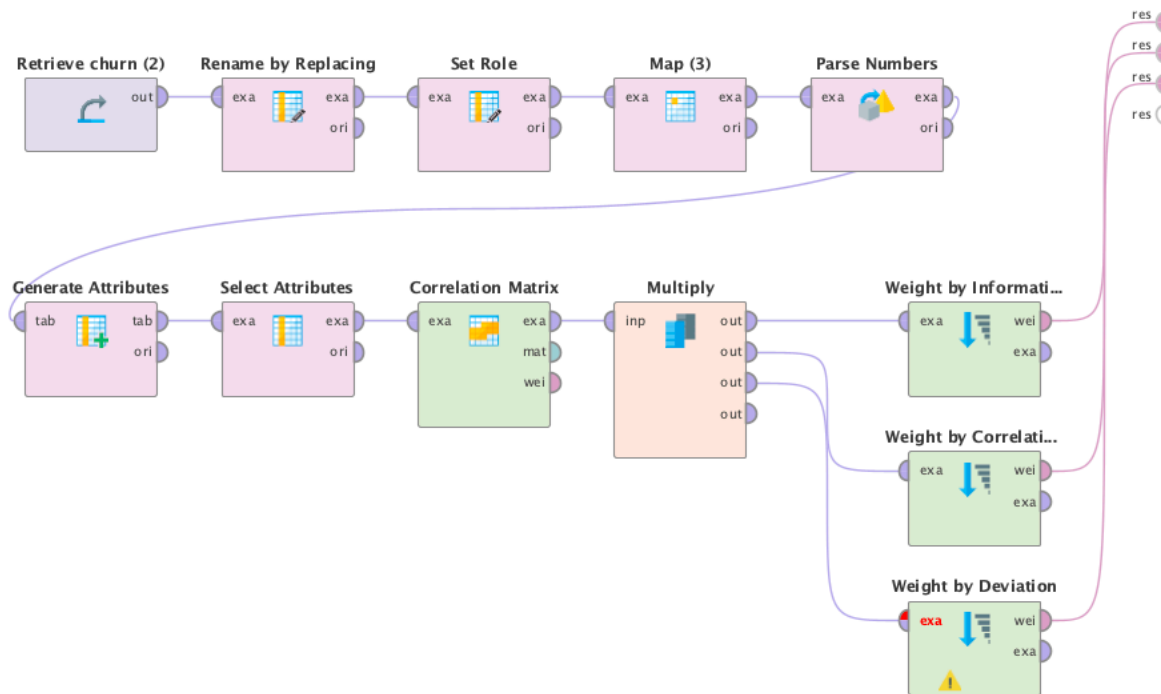
1.8. Tabla de variables

A continuación, se presenta la tabla de decisión sobre cada variable:

Variable	Decisión final
State	Eliminado. Los datos contienen errores
Account Length	Se mantiene
Area Code	Eliminado. Los datos contienen errores
Phone Number	Identificado como ID
Inter Plan	Se mantiene
VoiceMail Plan	Se elimina, pues tiene una correlación lineal alta con otra variable y fue fusionada con esta
No of Vmail Mesgs	Se elimina, pues tiene una correlación lineal alta con otra variable y fue fusionada con esta
Total Day Min	Se mantiene
Total Day Calls	Se mantiene
Total Day Charge	Se elimina, pues tiene una correlación lineal perfecta con otra variable
Total Evening Min	Se mantiene
Total Evening Calls	Se mantiene
Total Evening Charge	Se elimina, pues tiene una correlación lineal perfecta con otra variable
Total Night Min	Se mantiene
Total Night Calls	Se mantiene
Total Night Charge	Se elimina, pues tiene una correlación lineal perfecta con otra variable
Total Int Min	Se mantiene
Total Int Calls	Se mantiene
Total Int Charge	Se elimina, pues tiene una correlación lineal perfecta con otra variable
No of Calls Customer Service	Se mantiene
Churn	Variable objetivo
Total_Charge	Se crea esta nueva variable, dado que contiene información relevante
N_of_Vmail_Mesgs_VMail_Plan	Se crea esta nueva variable, dado que contiene información relevante

1.9. Operador Feature weights en RapidMiner

Para realizar selección de atributos con filtros se seleccionaron los operadores de peso basados en ganancia de información, desviación y correlación.



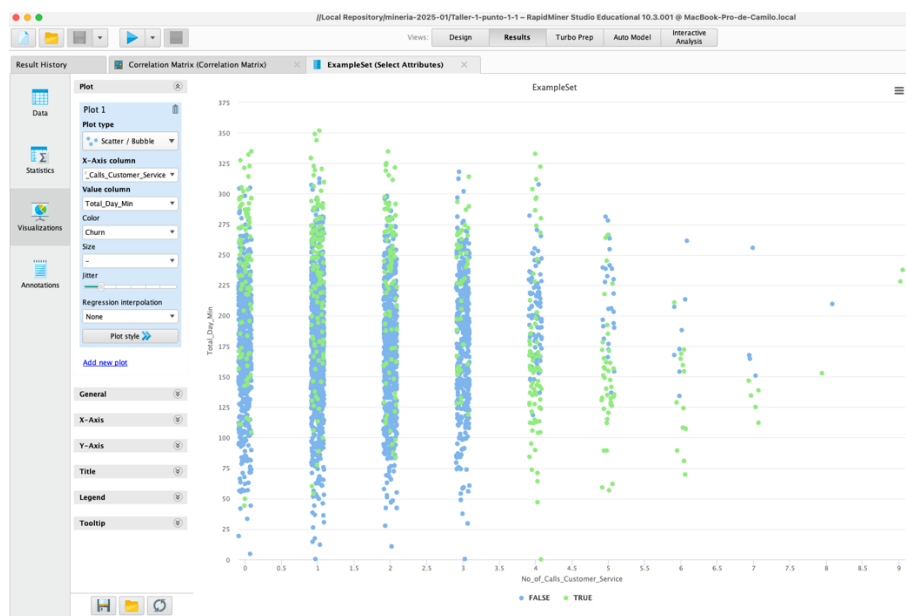
AttributeWeights (Weight by Deviation)		AttributeWeights (Weight by Information Gain)		AttributeWeights (Weight by Correlation)	
attribute	wei... ↓	attribute	wei... ↓	attribute	wei... ↓
Total_Day_Min	54.467	Total_Charge	0.129	Inter_Plan	0.260
Total_Evening_Min	50.714	Total_Day_Min	0.056	Total_Charge	0.232
Total_Night_Minutes	50.574	No_of_Calls_Customer_Service	0.050	No_of_Calls_Customer_Service	0.209
Account_Length	39.822	Inter_Plan	0.037	Total_Day_Min	0.205
Total_Day_calls	20.069	N_of_Vmail_Mesgs_VMail_Plan	0.008	Total_Evening_Min	0.093
Total_Evening_Calls	19.923	Total_Int_Min	0.007	N_of_Vmail_Mesgs_VMail_Plan	0.090
Total_Night_Calls	19.569	Total_Int_Calls	0.005	Total_Int_Min	0.068
N_of_Vmail_Mesgs_VMail_Plan	14.117	Total_Evening_Min	0.005	Total_Int_Calls	0.053
Total_Charge	10.502	Total_Night_Minutes	0.003	Total_Night_Minutes	0.035
Total_Int_Min	2.792	Total_Day_calls	0.002	Total_Day_calls	0.018
Total_Int_Calls	2.461	Total_Night_Calls	0.001	Account_Length	0.017
No_of_Calls_Customer_Service	1.315	Total_Evening_Calls	0.001	Total_Evening_Calls	0.009
		Account_Length	0.001	Total_Night_Calls	0.006

Al comparar los resultados obtenidos con los filtros con el análisis realizado, se puede observar que en los filtros de peso por ganancia de información y en el filtro de peso por correlación, los cinco primeros atributos priorizados son Total_charge, Total_day_min, Inter_plan, No_of_calls_customer_service y N_of_vmail_mesgs_vmail_plan, lo cual coincide con lo que se observó previamente.

Estos atributos presentan gran influencia sobre la decisión de un usuario de permanecer en la compañía telefónica o retirarse.

1.10.No of Calls Customer Service vs Total Day Minutes

En el grafico de puntos se puede observar que, de los clientes que realizaron entre cero y tres llamadas, los que más llamadas realizaban tenían una mayor cantidad de retiros del servicio, específicamente por encima de 200 llamadas. Por otra parte, después de 4 llamadas a servicio técnico el número de clientes que permaneció en la empresa bajo drásticamente, bajando su número de minutos consumidos en cada llamada al servicio técnico.



Realizando el cálculo del porcentaje de personas que permanecen y personas que se retiran en cada nueva llamada se obtiene la siguiente tabla:

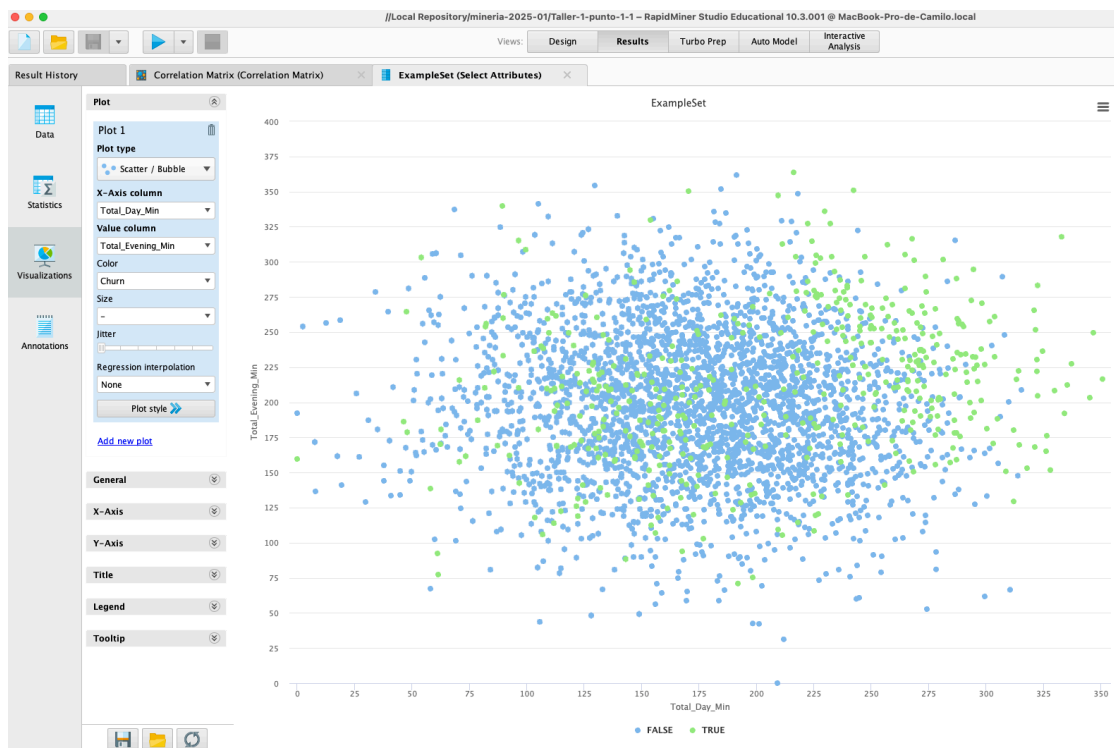
N of Calls	Churns	No Churns	% Churn	% No Churn
0	92	605	13,20%	86,80%
1	122	1059	10,33%	89,67%
2	87	672	11,46%	88,54%
3	44	385	10,26%	89,74%
4	76	90	45,78%	54,22%
5	40	26	60,61%	39,39%
6	14	8	63,64%	36,36%
7	5	4	55,56%	44,44%
8	1	1	50,00%	50,00%
9	2	0	100,00%	0,00%

Por encima de los 200 minutos de llamada el porcentaje de personas que cambian de compañía aumenta considerablemente entre 0 y tres llamadas, como se muestra a continuación:

N of Calls	Churns	No Churns	% Churn	% No Churn
0	65	204	24,16%	75,84%
1	87	323	21,22%	78,78%
2	59	183	24,38%	75,62%
3	29	148	16,38%	83,62%
4	14	46	23,33%	76,67%
5	5	16	23,81%	76,19%
6	1	3	25,00%	75,00%
7	0	1	0,00%	100,00%
8	0	1	0,00%	100,00%
9	2	0	100,00%	0,00%

1.11.Total Day Min vs Total Evening Min

Existe un área con incremento de personas que se retiran de la compañía cuando los minutos consumidos durante el día superan los 250 y los minutos en la tarde-noche superan los 175.



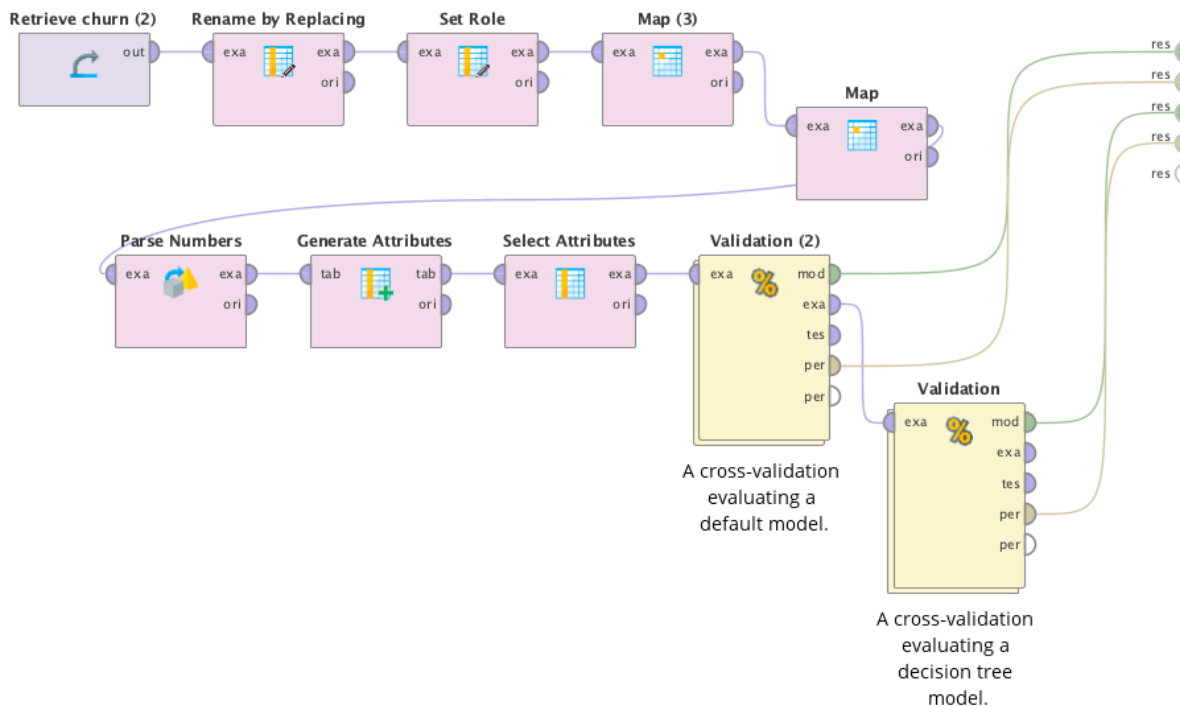
Realizando el cálculo del porcentaje de personas que permanecen y personas que se retiran dependiendo de numero de minutos en día y tarde-noche se obtiene la siguiente tabla:

Limite en Day Min	Limite en Evening Min	Churns	No Churns	% Churn	% No Churn
250	175	137	104	56,85%	43,15%

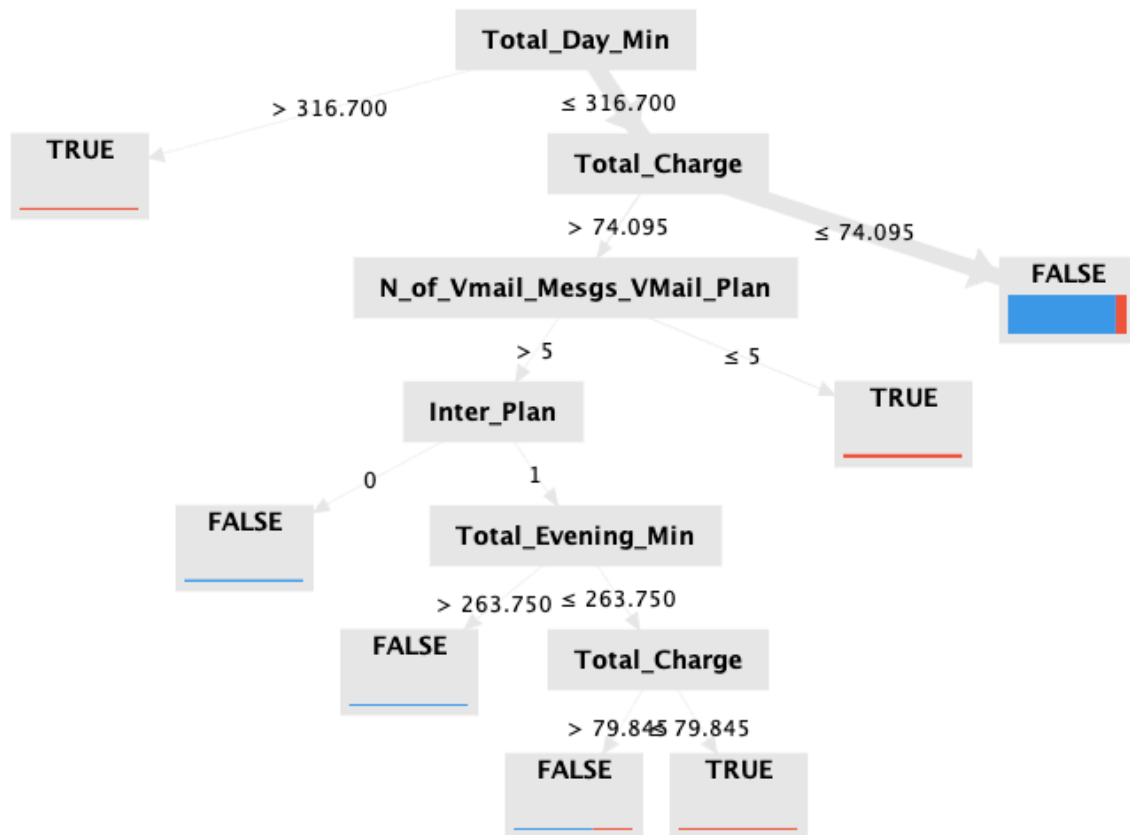
Como se puede apreciar, esta es un área de gran concentración de personas que se retiran, siendo más de la mitad de los usuarios.

2. Clasificador: Decision Trees

Se crea un árbol de decisión como modelo de clasificación utilizado un bloque de validación cruzada nominal, utilizando la semilla 1992.



El árbol generado se muestra a continuación:



2.1. Conclusiones al observar el árbol

Como se pudo apreciar en el punto 1.7 en la gráfica de cargo total y personas que cambian de compañía telefónica, si este valor es superior a 74 los clientes tienden a cambiar de compañía, demostrando que este valor es de gran importancia para los clientes. Como se preveía en el análisis del punto 1, las cuatro variables (Total charge, Total day min, Total evening min e Inter plan) son determinantes para tomar una decisión en la predicción realizada. Por otro lado, el número de mensajes de voz y el plan de voz son importantes para la decisión del usuario de abandonar la compañía una vez se han cumplido las condiciones que se observan en el árbol de decisión, lo cual tiene relación con el análisis realizado.

2.2. Evaluación de performance

PerformanceVector (Performance (2))			
Criterion	Table View		
accuracy			
precision			
recall			
AUC (optimistic)			
AUC			
AUC (pessimistic)			

accuracy: 91.30% +/- 1.08% (micro average: 91.30%)

	true FALSE	true TRUE	class precision
pred. FALSE	2846	286	90.87%
pred. TRUE	4	197	98.01%
class recall	99.86%	40.79%	

PerformanceVector (Performance (2))			
Criterion	Table View		
accuracy			
precision			
recall			
AUC (optimistic)			
AUC			
AUC (pessimistic)			

precision: 98.00% +/- 2.63% (micro average: 98.01%) (positive class: TRUE)

	true FALSE	true TRUE	class precision
pred. FALSE	2846	286	90.87%
pred. TRUE	4	197	98.01%
class recall	99.86%	40.79%	

PerformanceVector (Performance (2))			
Criterion	Table View		
accuracy			
precision			
recall			
AUC (optimistic)			
AUC			
AUC (pessimistic)			

recall: 40.81% +/- 7.18% (micro average: 40.79%) (positive class: TRUE)

	true FALSE	true TRUE	class precision
pred. FALSE	2846	286	90.87%
pred. TRUE	4	197	98.01%
class recall	99.86%	40.79%	

Los valores predichos como falsos que realmente son verdaderos son los datos que se deberían optimizar, ya que son errores que pueden afectar en mayor medida a la empresa, ya que son personas que se están retirando sin que la empresa lo sepa. Se necesitaría una medida como la exactitud que tomara en cuenta que los verdaderos negativos no aumenten, ya que esto afecta directamente al negocio.

2.3. Comparativa con el modelo por defecto

A continuación, se muestra el performance observando la exactitud con el clasificador por defecto utilizando la semilla 1992:

PerformanceVector (Performance)			
Criterion	Table View		
accuracy			
precision			
recall			
AUC (optimistic)			
AUC			
AUC (pessimistic)			

accuracy: 85.51% +/- 0.12% (micro average: 85.51%)

	true FALSE	true TRUE	class precision
pred. FALSE	2850	483	85.51%
pred. TRUE	0	0	0.00%
class recall	100.00%	0.00%	

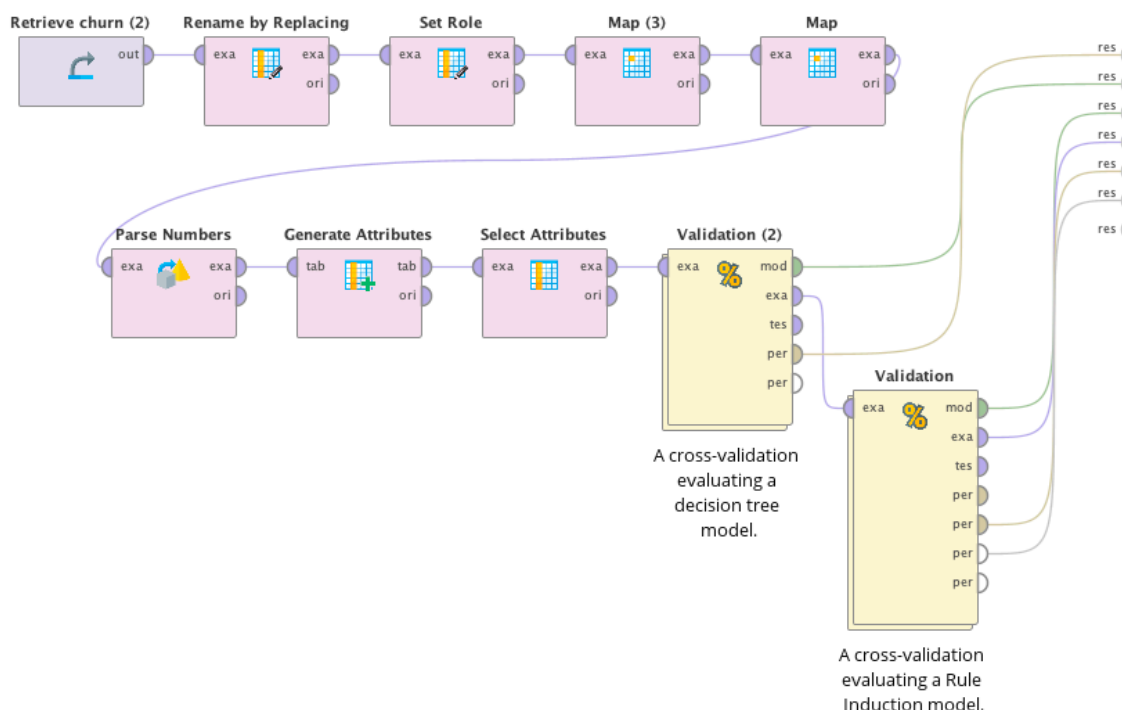
El modelo por defecto asume que el valor de Churn será falso para todos los usuarios, siendo una predicción sin valor para la empresa. Como se puede apreciar, el exactitud del modelo por defecto es inferior al que se obtiene con el árbol de decisión, como era esperable.

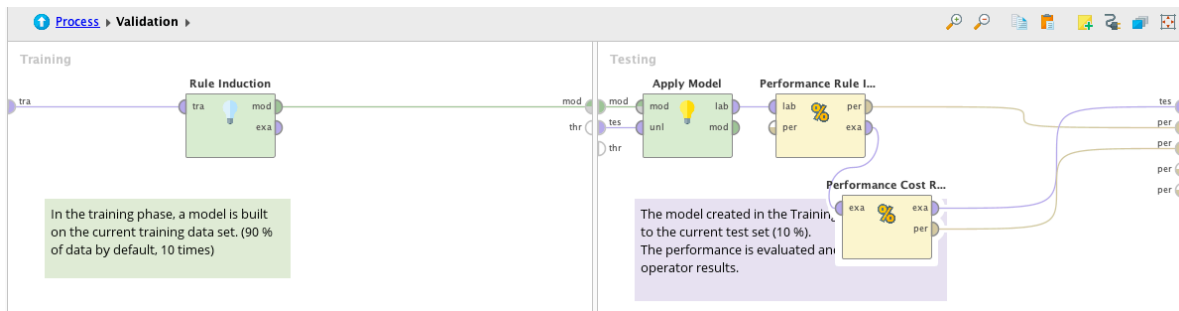
Tree (Decision Tree)			
Result History			
Criterion	Table View <input checked="" type="radio"/> Plot View <input type="radio"/>		
accuracy			
precision			
recall	recall: 0.00% +/- 0.00% (micro average: 0.00%) (positive class: TRUE)		
AUC (optimistic)			
AUC			
AUC (pessimistic)			
	true FALSE	true TRUE	class precision
pred. FALSE	2850	483	85.51%
pred. TRUE	0	0	0.00%
class recall	100.00%	0.00%	

Adicionalmente, el recall del modelo es de 0%, dando a entender que es un modelo que permite pasar todos los errores de clasificación que se presenten.

3. Clasificador Rules Induction

Tomando en cuenta el análisis realizado y los resultados obtenidos hasta el momento, se eliminan las variables: “Total Night Min”, “Total Night Calls” y “Account Length”. Al hacer esto con el árbol de decisión se obtuvo una exactitud de 92.02, una precisión de 97.69 y un recall de 46.02, mejorando con respecto al modelo anterior en la exactitud y el recall.





3.1. Conclusiones del nuevo modelo

A continuación, se presentan las reglas generadas por el operador “Rule Induction”, en las cuales se tienen en cuenta las variables “Total Charge” y “N of Vmail Mesgs Vmail Plan”, las cuales fueron las variables que se crearon en el punto 1.2. Estas reglas son mucho más sencillas que el árbol de decisión y logran un desempeño similar.

RuleModel (Rule Induction)

Result History

PerformanceVector (Performance Rule Induction)

Description

if Total_Charge ≤ 74.035 then FALSE (2782 / 276)
 if N_of_Vmail_Mesgs_VMail_Plan ≤ 5 then TRUE (0 / 200)
 else FALSE (61 / 6)

Annotations

correct: 3043 out of 3325 training examples.

3.2. Performance measures

History

Criterion

accuracy

precision

recall

AUC (optimistic)

AUC

AUC (pessimistic)

PerformanceVector (Performance Rule Induction)

Table View Plot View

accuracy: 91.45% +/- 1.02% (micro average: 91.45%)

	true FALSE	true TRUE	class precision
pred. FALSE	2849	284	90.94%
pred. TRUE	1	199	99.50%
class recall	99.96%	41.20%	

PerformanceVector (Performance Rule Induction)			
Table View Plot View			
precision: 99.41% +/- 1.86% (micro average: 99.50%) (positive class: TRUE)			
	true FALSE	true TRUE	class precision
pred. FALSE	2849	284	90.94%
pred. TRUE	1	199	99.50%
class recall	99.96%	41.20%	

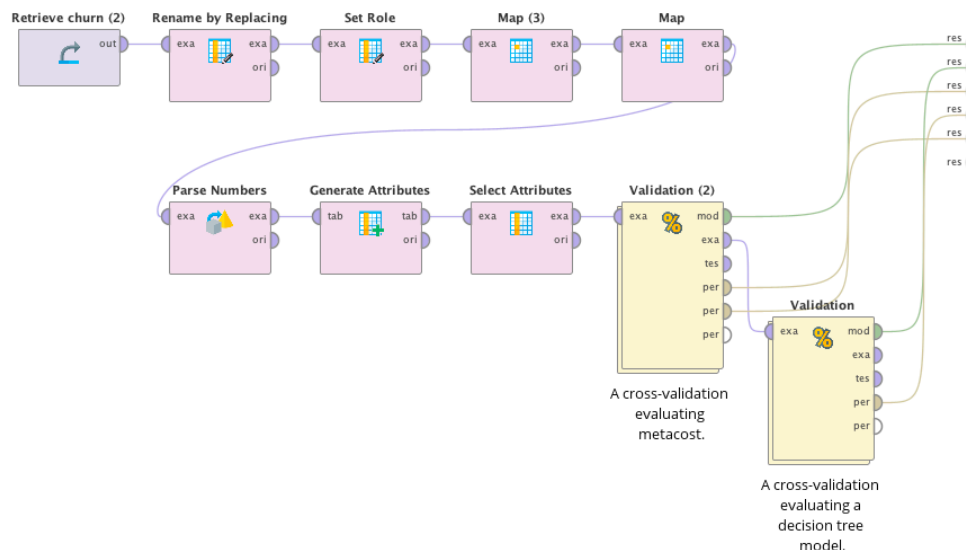
PerformanceVector (Performance Rule Induction)			
Table View Plot View			
recall: 41.22% +/- 6.57% (micro average: 41.20%) (positive class: TRUE)			
	true FALSE	true TRUE	class precision
pred. FALSE	2849	284	90.94%
pred. TRUE	1	199	99.50%
class recall	99.96%	41.20%	

Como se puede apreciar, el modelo obtiene buenos resultados, siendo muy cercanos a los obtenidos por el árbol de decisión.

3.3. Comparativa con el modelo por defecto

La exactitud obtenida por el modelo por defecto es de 85.51%, siendo mucho menor a la exactitud de. 91.45% obtenida con el modelo de reglas, como podría esperarse teniendo en cuenta que el modelo por defecto asume que nadie se va de la compañía. Adicionalmente, el modelo por defecto tiene un 0% de recall, lo cual no es útil para una clasificación.

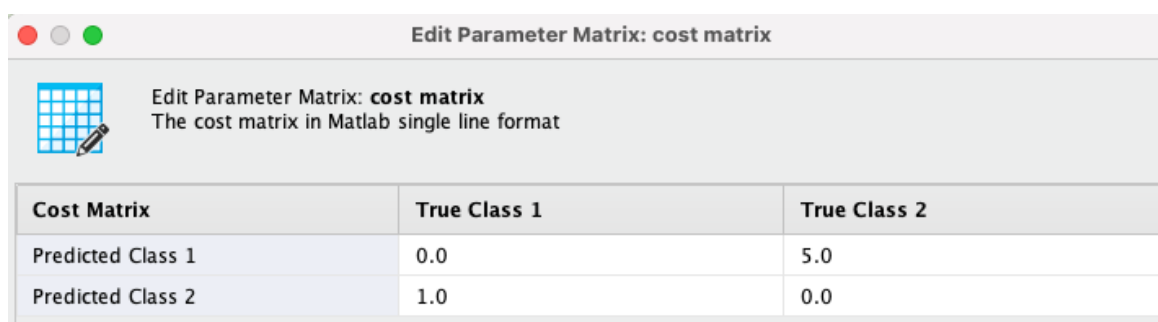
4. Modelo sensible al costo



4.1. Datos desde la perspectiva de los costos

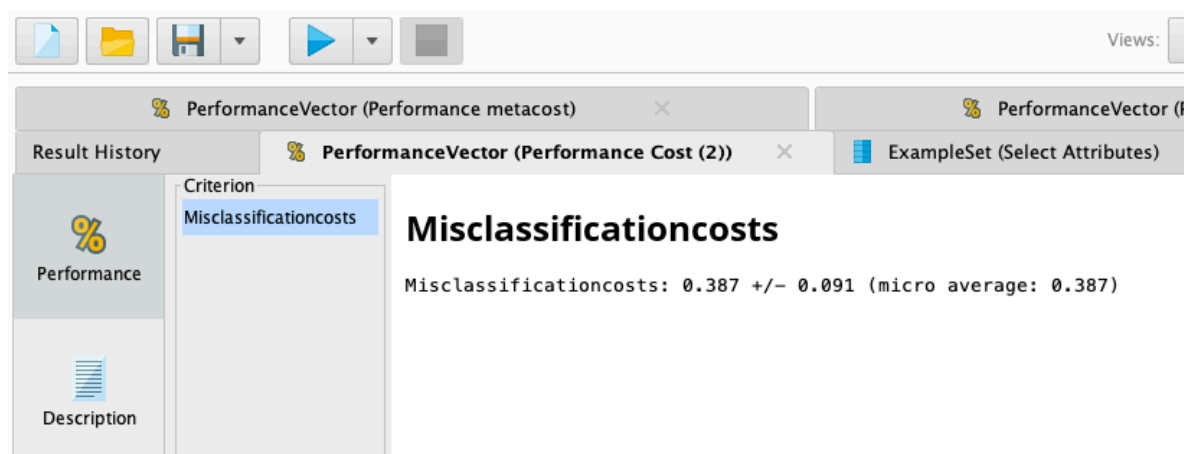
El costo de una mala decisión puede ser alto y no se está teniendo en cuenta hasta el momento, para demostrarlo se plantea una matriz de costo donde el valor de una predicción de churn falso que en realidad fuese verdadero tenga mayor peso al caso contrario (churn predicho verdadero, churn real falso).

El costo de perder a un cliente que está dentro de la compañía implica volver a captar la atención del cliente, ya sea con ofertas que reduzcan el pago de su factura por un determinado plazo. Además de esto, la persona perjudica el mercado ya que no recomendará a la compañía y hablara mal de esta. Por estos motivos se asumirá que la relación de que una persona que se predijo que se iría en realidad se quede, con respecto a una persona que se predecía que se quedaría pero en realidad se va, será de 1 a 5.



Cost Matrix	True Class 1	True Class 2
Predicted Class 1	0.0	5.0
Predicted Class 2	1.0	0.0

Utilizando un operador de performance de costo se puede observar que el costo de la decisión es de 0.387.



Views: [icon] [icon] [icon] [icon] [icon]

PerformanceVector (Performance metacost) [X] PerformanceVector (P [X]

Result History [icon] PerformanceVector (Performance Cost (2)) [X] ExampleSet (Select Attributes)

Criterion: Misclassificationcosts

Misclassificationcosts

Misclassificationcosts: 0.387 +/- 0.091 (micro average: 0.387)

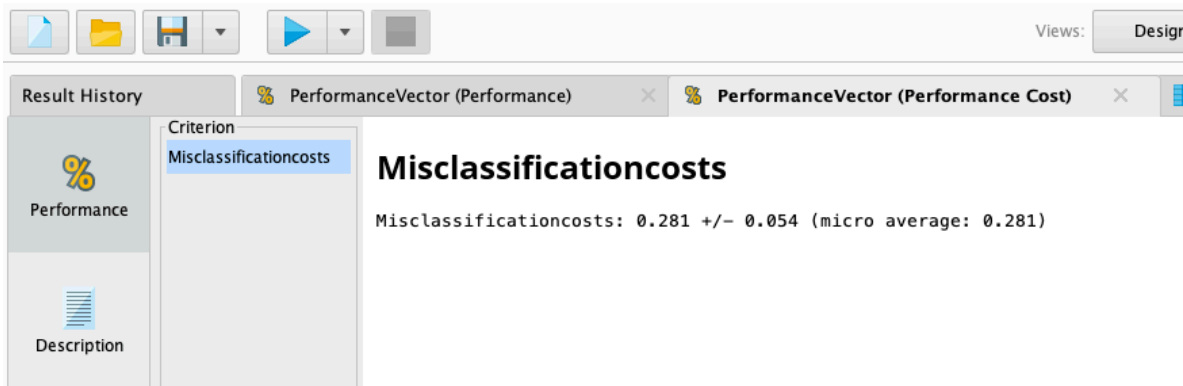
Performance [icon]

Description [icon]

4.2. Solución al problema

Utilizando el operador metacost se puede guiar la decisión que tome un árbol de decisión, buscando disminuir este costo en la predicción que realice el modelo.

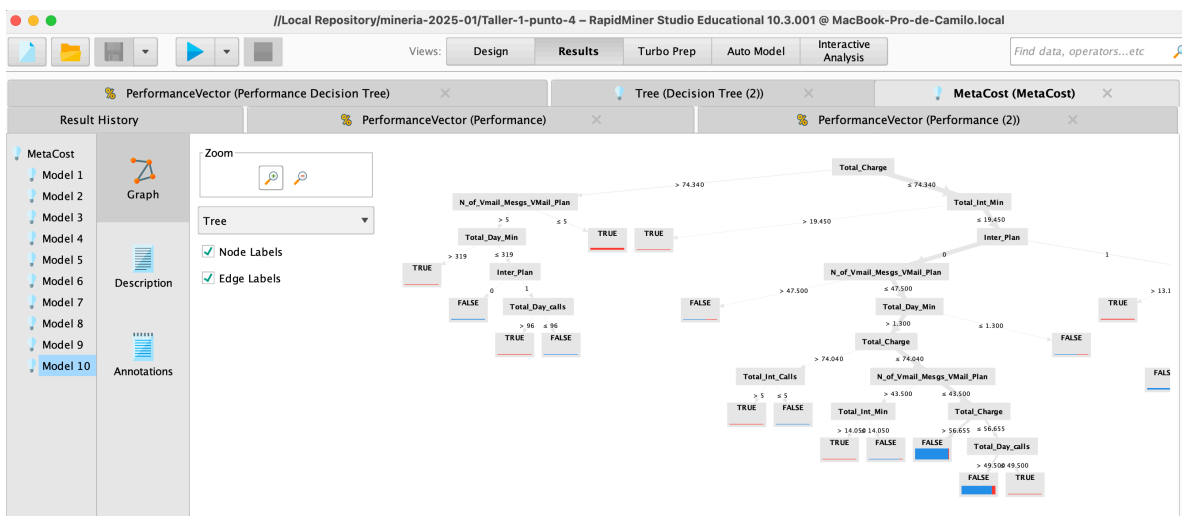
Utilizando un operador de performance que obtenga el costo de la decisión se obtiene el siguiente resultado:



Al buscar disminuir el costo de la decisión, el desempeño obtenido fue de 0.281.

4.3. Modelo obtenido

El operador metacost tiene un comportamiento de ensamble tipo bagging, en el cual se crean múltiples modelos (árboles de decisión) que permiten realizar la predicción. Uno de estos modelos se muestra a continuación:



Los árboles generados tienden a ser más complejos que el árbol creado por el decision tree pero mantienen las mismas variables.

4.4. Comparativa de desempeño

El desempeño del modelo basado en costos se muestra a continuación:

ExampleSet (Select Attributes) Tree (Decision Tree (2)) PerformanceVector (Performance metacost) PerformanceVector (Perf

Criterion
accuracy
precision
recall
AUC (optimistic)
AUC
AUC (pessimistic)

Table View Plot View

accuracy: 93.25% +/- 1.17% (micro average: 93.25%)

	true FALSE	true TRUE	class precision
pred. FALSE	2803	178	94.03%
pred. TRUE	47	305	86.65%
class recall	98.35%	63.15%	

ExampleSet (Select Attributes) Tree (Decision Tree (2)) PerformanceVector (Performance metacost) PerformanceVector (Perf

Criterion
accuracy
precision
recall
AUC (optimistic)
AUC
AUC (pessimistic)

Table View Plot View

precision: 86.86% +/- 5.24% (micro average: 86.65%) (positive class: TRUE)

	true FALSE	true TRUE	class precision
pred. FALSE	2803	178	94.03%
pred. TRUE	47	305	86.65%
class recall	98.35%	63.15%	

ExampleSet (Select Attributes) Tree (Decision Tree (2)) PerformanceVector (Performance metacost) PerformanceVector (Perf

Criterion
accuracy
precision
recall
AUC (optimistic)
AUC
AUC (pessimistic)

Table View Plot View

recall: 63.15% +/- 7.50% (micro average: 63.15%) (positive class: TRUE)

	true FALSE	true TRUE	class precision
pred. FALSE	2803	178	94.03%
pred. TRUE	47	305	86.65%
class recall	98.35%	63.15%	

Performance	Decision tree	Rule induction	Metacost decision tree
Accuracy	92.02%	91.45%	93.25%
Precision	97.69%	99.41%	86.86%
Recall	46.02%	41.22%	63.15%
Costo	0.387	0.426	0.281

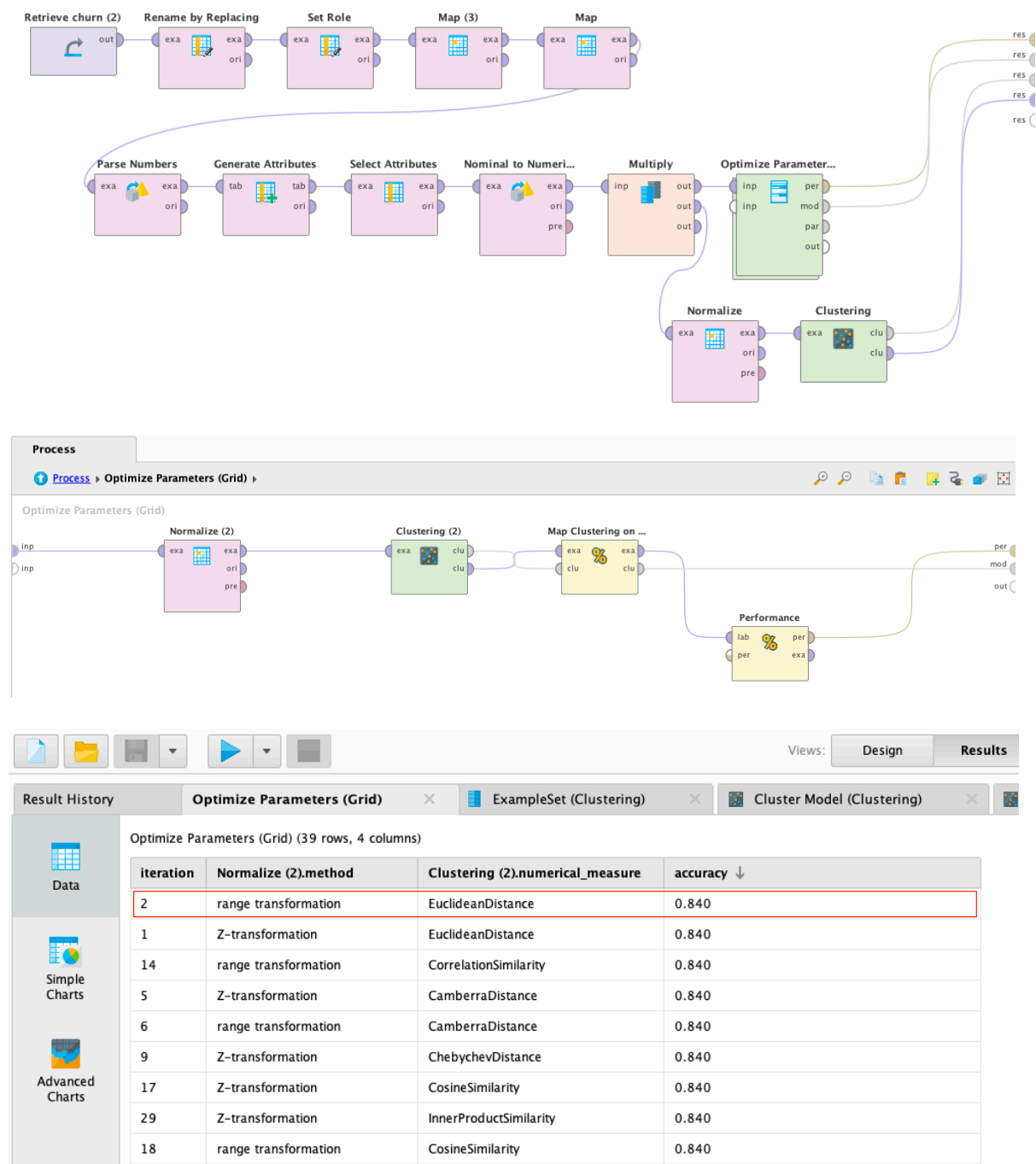
Como se puede observar, el modelo obtenido priorizando el costo de la decision obtiene un mejor recall al compararlo con los otros dos modelos obtenidos previamente, así como una mejor exactitud. Sin embargo, su precisión baja considerablemente, siendo 11% menos que la obtenida por el decision tree y 13% menos que la obtenida por el modelo rule induction. Observando el cálculo del costo, se obtuvo un valor mucho más alto en el modelo Rule induction, un valor intermedio con el modelo decisión tree, y un valor mucho menor con el ensamble del metacost, además se puede apreciar que los valores “True False” que son de mayor interés se redujeron, lo cual era el resultado esperado.

5. Clustering

En este punto se busca obtener dos agrupaciones, True Churn y False Churn que se asemejen lo más posible al valor real con el que se cuenta. Se simula entonces

un ejercicio con datos no supervisados en el que no se tendrá en cuenta la variable objetivo para guiar la clasificación.

Utilizando el bloque de optimización de parámetros se busca cual será el método de normalización y la técnica de clustering que mejor exactitud ofrece.



Como se puede observar, con una transformación por rango, junto con la distancia Euclidea obtienen la mejor exactitud con el menor número de iteraciones.

5.1. Descripción de los clusters

Como se menciono anteriormente, se escogen dos cluster de tal manera que predigan el comportamiento de la variable churn, la cual no está guiando la clasificación para este ejercicio. Los resultados obtenidos por el operador de cluster se muestran a continuación:

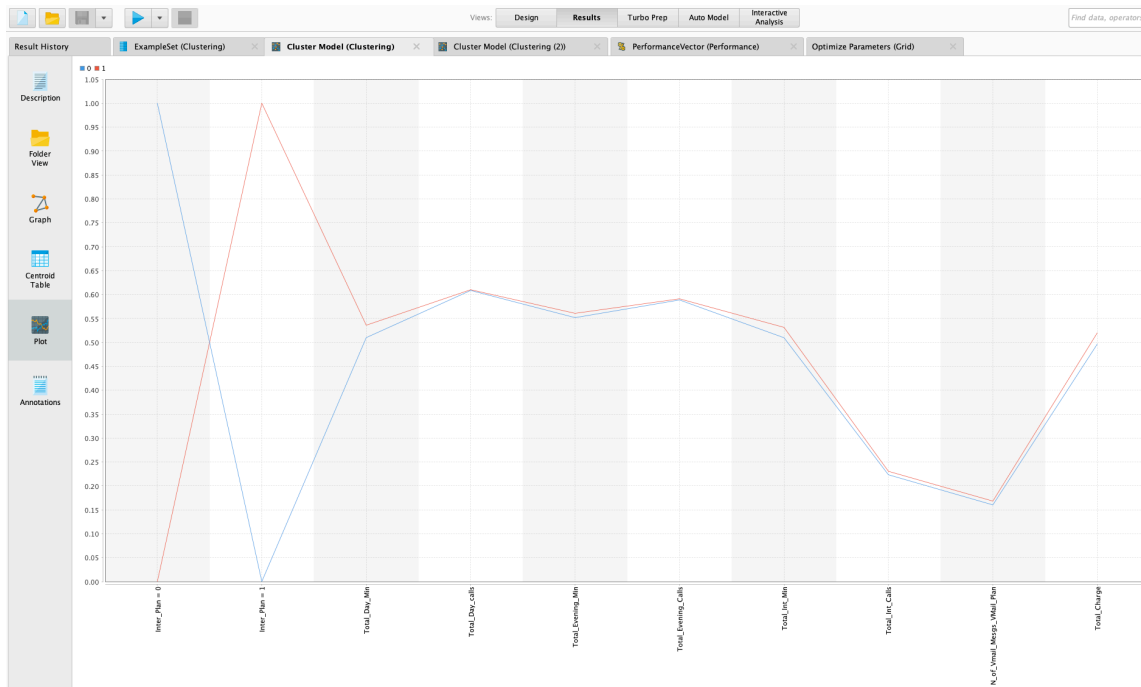
Row No.	Phone_Nu...	Churn ↑	cluster	Inter_Plan ...	Inter_Plan ...	Total_Day_...	Total_Day_...	Total_Eveni...	Total_Eveni...	Total_Int_...	Total_Int_C...	N_of_Vmail...	Total_Char...
1	371-7191	FALSE	cluster_0	1	0	0.461	0.745	0.538	0.606	0.685	0.150	0.519	0.496
2	358-1921	FALSE	cluster_0	1	0	0.694	0.691	0.333	0.647	0.610	0.250	0	0.538
3	375-9999	FALSE	cluster_1	0	1	0.853	0.430	0.170	0.518	0.330	0.350	0	0.599
4	330-6626	FALSE	cluster_1	0	1	0.475	0.685	0.408	0.718	0.505	0.150	0	0.398
5	391-8027	FALSE	cluster_1	0	1	0.637	0.594	0.607	0.594	0.315	0.300	0	0.610
6	355-9993	FALSE	cluster_0	1	0	0.622	0.533	0.958	0.635	0.375	0.350	0.481	0.756
7	329-9001	FALSE	cluster_1	0	1	0.448	0.479	0.283	0.553	0.355	0.300	0	0.327
8	335-4719	FALSE	cluster_0	1	0	0.526	0.588	0.967	0.471	0.435	0.200	0	0.688
9	330-8173	FALSE	cluster_1	0	1	0.737	0.509	0.610	0.653	0.560	0.250	0.731	0.787
11	344-9403	FALSE	cluster_0	1	0	0.535	0.770	0.449	0.871	0.455	0.250	0	0.466
12	363-1107	FALSE	cluster_0	1	0	0.367	0.582	0.288	0.418	0.560	0.100	0	0.236
13	394-8006	FALSE	cluster_0	1	0	0.446	0.533	0.681	0.441	0.615	0.250	0	0.501
14	366-9238	FALSE	cluster_0	1	0	0.344	0.424	0.845	0.447	0.655	0.300	0	0.497

	true FALSE	true TRUE	class precision
pred. FALSE	2664	346	88.50%
pred. TRUE	186	137	42.41%
class recall	93.47%	28.36%	

Se puede observar que la exactitud del modelo bajo considerablemente.

5.2. Información útil para el problema de clasificación

Gracias a los clusters obtenidos es posible observar la relación que existe entre las variables de interés y la agrupación realizada.



Para el algoritmo de clustering la diferencia principal de un cluster a otro se basa en el atributo Inter Plan, junto con una pequeña diferenciación entre las variables Total Day Min, Total Int Min y Total Charge. Esto puede afirmarse con una exactitud del 84.04%.