Diplomado Data Science: Desafío de clasificación

Alumno: Roxana Cares

1. Contexto del problema:

Una empresa de telecomunicaciones requiere implementar un modelo de clasificación que permita identificar clientes que se van a fugar, de manera que puedan gestionar algún mecanismo de retención/fidelización. La empresa proporcionó un conjunto de datos de datos (churn-analysis.csv) con información que puede ser de utilidad para identificar patrones de comportamiento de clientes que se van a fugar.

2. Análisis Exploratorio de los Datos:

Con este set de datos se realizó, en primer lugar, un análisis exploratorio con el fin de familiarizarse con los datos y revisar su naturaleza.

En la figura 1 podemos observar que el set de datos cuenta con 20 variables en total, donde la variable a predecir corresponde a la fuga de clientes (“churn”) y está representada con los valores “True”, “False”. Esta es una variable que posee un evidente desbalance entre sus dos clases, en donde la modalidad “False” se encuentra sobrerrepresentada (alrededor del 85% del total de datos).

```
     state         area.code     phone.number  international.plan voice.mail.plan number.vmail.messages total.day.minutes
 WV     : 106   408: 838    327-1058:   1   no :3010           no :2411        Min.   : 0.000        Min.   :  0.0
 MN     :  84   415:1655    327-1319:   1   yes: 323           yes: 922        1st Qu.: 0.000        1st Qu.:143.7
 NY     :  83   510: 840    327-3053:   1                                     Median : 0.000        Median :179.4
 AL     :  80               327-3587:   1                                     Mean   : 8.099        Mean   :179.8
 OH     :  78               327-3850:   1                                     3rd Qu.:20.000        3rd Qu.:216.4
 OR     :  78               327-3954:   1                                     Max.   :51.000        Max.   :350.8
 (Other):2824               (Other) :3327
 total.day.calls total.day.charge total.eve.minutes total.eve.calls total.eve.charge total.night.minutes total.night.calls
 Min.   :  0.0   Min.   : 0.00    Min.   :  0.0     Min.   :  0.0   Min.   : 0.00    Min.   : 23.2       Min.   : 33.0
 1st Qu.: 87.0   1st Qu.:24.43    1st Qu.:166.6     1st Qu.: 87.0   1st Qu.:14.16    1st Qu.:167.0       1st Qu.: 87.0
 Median :101.0   Median :30.50    Median :201.4     Median :100.0   Median :17.12    Median :201.2       Median :100.0
 Mean   :100.4   Mean   :30.56    Mean   :201.0     Mean   :100.1   Mean   :17.08    Mean   :200.9       Mean   :100.1
 3rd Qu.:114.0   3rd Qu.:36.79    3rd Qu.:235.3     3rd Qu.:114.0   3rd Qu.:20.00    3rd Qu.:235.3       3rd Qu.:113.0
 Max.   :165.0   Max.   :59.64    Max.   :363.7     Max.   :170.0   Max.   :30.91    Max.   :395.0       Max.   :175.0

 total.night.charge total.intl.minutes total.intl.calls total.intl.charge customer.service.calls   churn
 Min.   : 1.040     Min.   : 0.00      Min.   : 0.000   Min.   :0.000     Min.   :0.000          False:2850
 1st Qu.: 7.520     1st Qu.: 8.50      1st Qu.: 3.000   1st Qu.:2.300     1st Qu.:1.000          True : 483
 Median : 9.050     Median :10.30      Median : 4.000   Median :2.780     Median :1.000
 Mean   : 9.039     Mean   :10.24      Mean   : 4.479   Mean   :2.765     Mean   :1.563
 3rd Qu.:10.590     3rd Qu.:12.10      3rd Qu.: 6.000   3rd Qu.:3.270     3rd Qu.:2.000
 Max.   :17.770     Max.   :20.00      Max.   :20.000   Max.   :5.400     Max.   :9.000
```

Figura 1. Resumen descriptivo de las variables del set de datos.

Por otro lado, el código de área suele estar relacionado con el estado. Dado que la variable state tiene 51 modalidades y ninguna de ellas presenta una mayor representatividad que las demás (Figura 2), se decidió mantener area.code como representante de la ubicación geográfica. Esta elección se basa en que area.code tiene solo 3 modalidades, las cuales están más balanceadas entre sí.

Finalmente, phone.number es una variable que no aporta información en el contexto de este estudio pues es único para cada cliente.

En resumen, contamos con 14 variables numéricas y 3 variables categóricas como posibles predictoras de la fuga de clientes.

state	n	porcentaje
<fct>	<int>	<dbl>
1 WV	106	3.18
2 MN	84	2.52
3 NY	83	2.49
4 AL	80	2.40
5 OH	78	2.34
6 OR	78	2.34
7 WI	78	2.34
8 VA	77	2.31
9 WY	77	2.31
10 CT	74	2.22

i 41 more rows

area.code	n	porcentaje
<fct>	<int>	<dbl>
1 415	1655	49.7
2 510	840	25.2
3 408	838	25.1

> |

Figura 2. Representación de las variables state y area.code en el set de datos.

Respecto a las otras variables categóricas, podemos observar un desbalance de las modalidades. El 90% de los clientes no cuenta con un plan internacional y cerca del 72% no posee correo de voz (Figura 1 y 4).

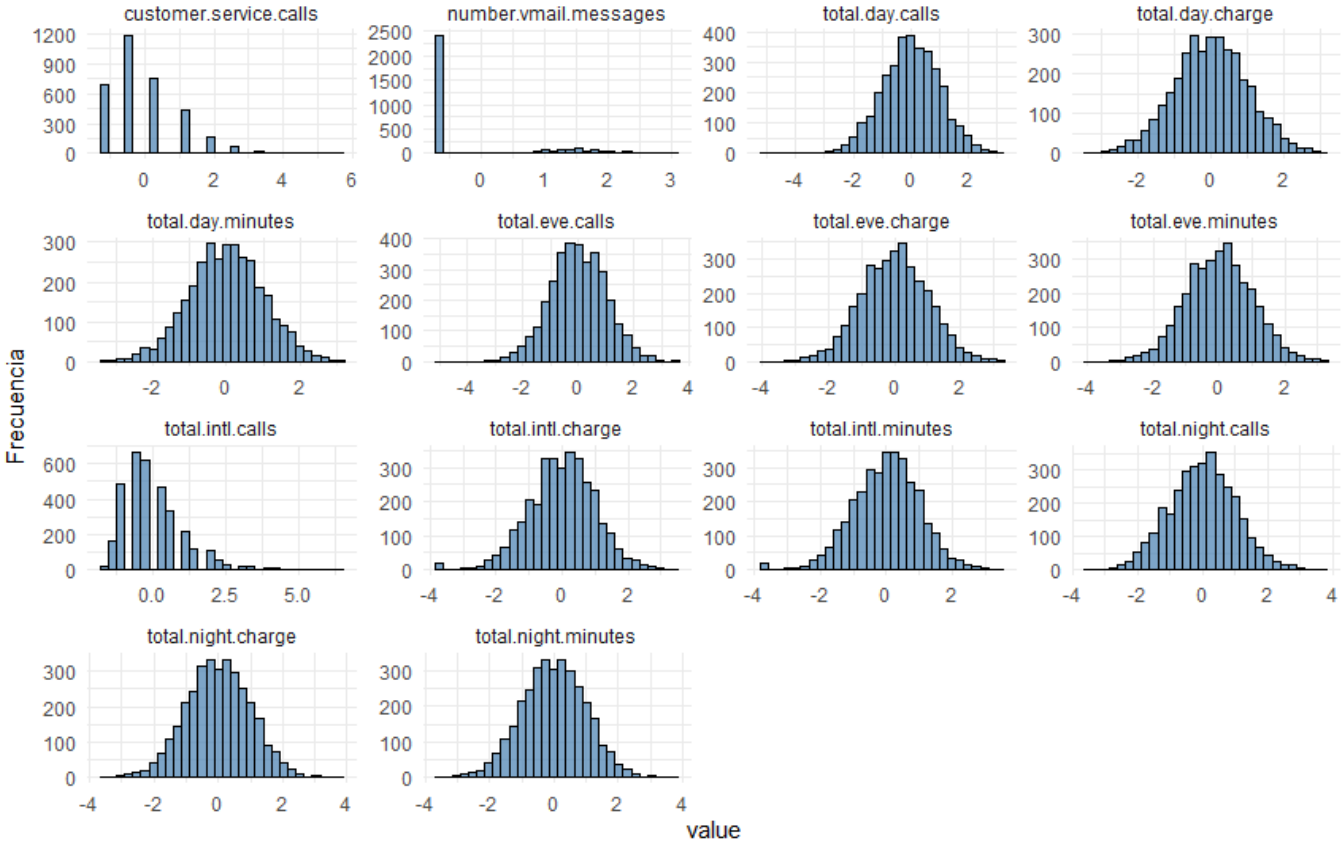


Figura 3. Gráficas de distribución de frecuencias de las variables numéricas del set de datos.

Con el fin de visualizar y comparar las variables numéricas en una misma escala, se realizó una transformación que las estandarizó a media 0 y desviación estándar 1. En la Figura 2, se puede observar que la mayoría de las variables tiende a una distribución normal, lo cual concuerda con la similitud entre las medias y las medianas en estos casos (Figura 1). Sin embargo, hay tres excepciones: la cantidad de llamadas a la mesa de ayuda (customer.service.calls), la cantidad de mensajes que posee el usuario (number.vmail.messages) y la cantidad de llamadas internacionales (total.intl.calls).

Estas variables tienden a concentrarse en valores del extremo inferior de la distribución y podemos observar específicamente que la mayoría de los datos del número de mensajes se agrupan en torno a un valor específico (cero, Figura 1), mientras que unos pocos datos se encuentran desplazados hacia valores más altos en el extremo izquierdo. Si cruzamos esta información con la de las variables categóricas, no es de extrañar que la mayoría de los usuarios registre cero mensajes si, en realidad, el 72% no posee correo de voz.

Al analizar la Figura 5, observamos que el grupo de clientes que no se fugan (churn “False”) presenta más datos atípicos en general en las variables, evidenciados por los puntos que caen fuera de los bigotes del gráfico. Esto podría indicar que estos clientes son más propensos a exhibir comportamientos extremos.

Por otro lado, en el grupo de clientes que se han fugado (churn “True”), se evidencia una mayor variabilidad en las variables relacionadas con las llamadas diarias, el costo diario y las interacciones con el servicio al cliente. Esta variabilidad sugiere que los clientes que se fugan podrían tener patrones de comportamiento más diversos en comparación con aquellos que permanecen. Sin embargo, es crucial tener en cuenta que la menor cantidad de datos en este grupo puede limitar la generalización de estos hallazgos.

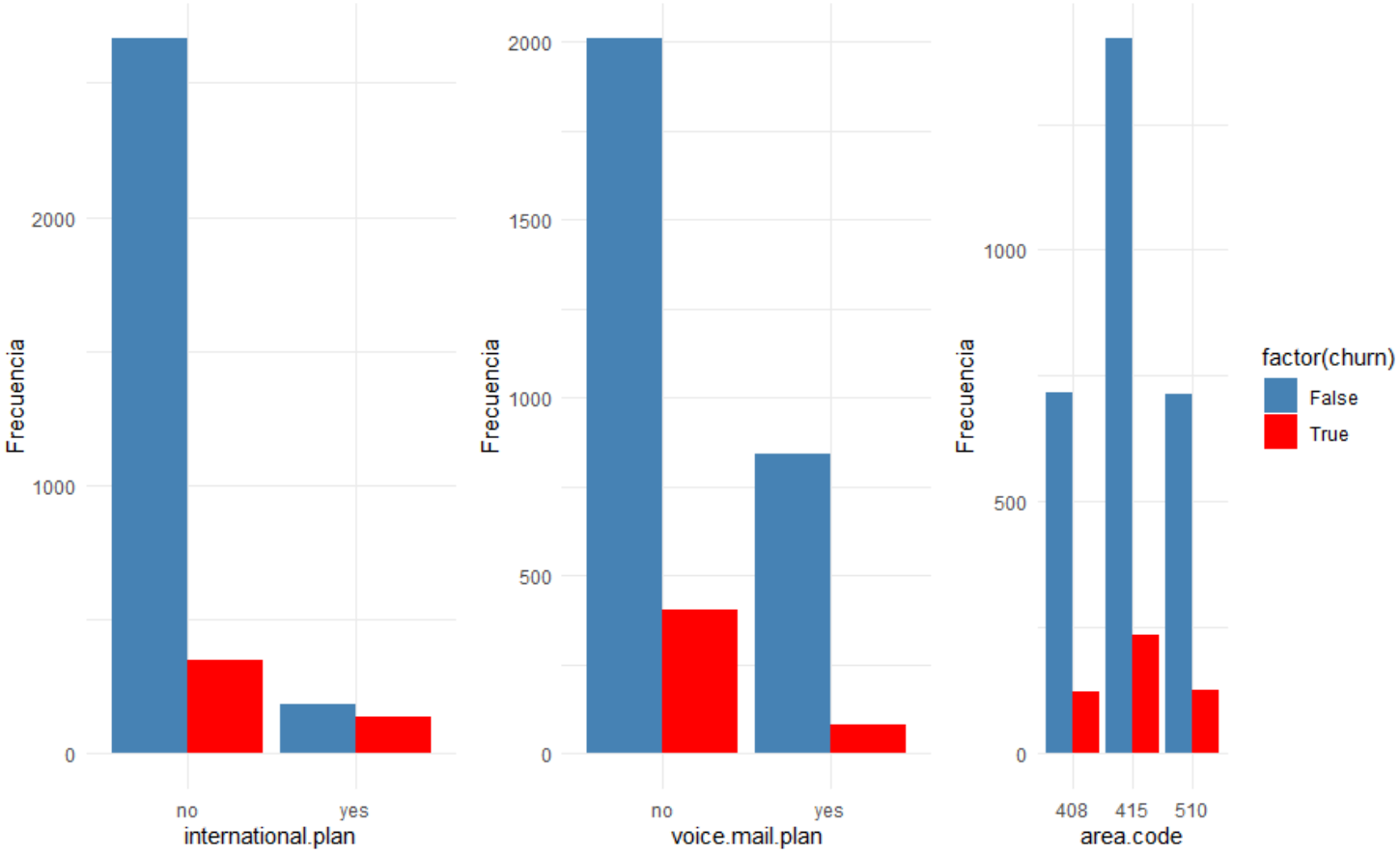


Figura 4. Gráfico de frecuencia para las variables categóricas del set de datos de acuerdo a la fuga o no de clientes.

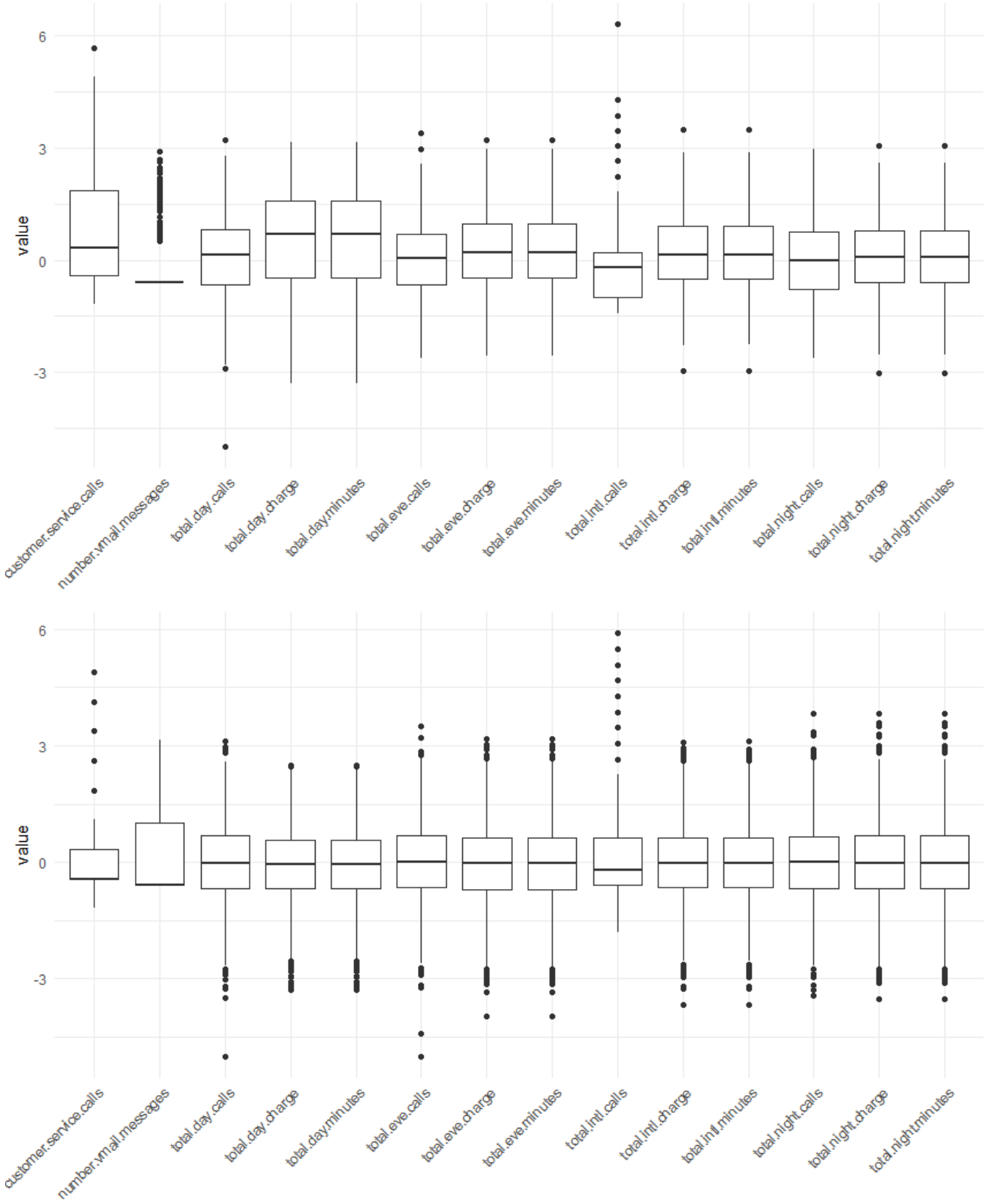


Figura 5. Gráfica de boxplot de las variables numéricas del set de datos para la categoría churn true (arriba) y false (abajo).

3. **Tratamiento de Datos Faltantes y Atípicos**:

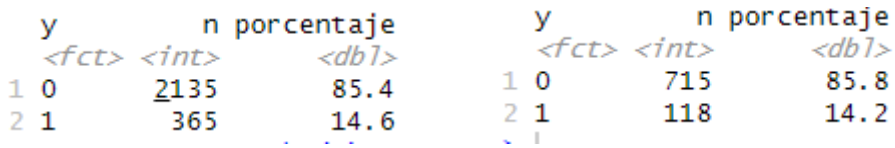
El conjunto de datos no presenta valores nulos.

En la sección anterior se observó que tres variables presentan distribuciones inusuales, con una dispersión de datos que sugiere la presencia de valores atípicos. En general, al analizar la Figura 5, se puede ver que casi todas las variables incluyen datos fuera del rango intercuartil ±1.5 unidades, los cuales podrían considerarse outliers. Sin embargo, estos datos atípicos no parecen ser errores de medición y podrían contener información relevante respecto a la fuga de clientes. Por lo tanto, no eliminaremos ningún dato en esta etapa, sino que primero evaluaremos el rendimiento del modelo de predicción.

Dado que la mayoría de los usuarios registra cero mensajes (Figura 1), este comportamiento podría considerarse normal, mientras que aquellos pocos usuarios con mensajes podrían calificarse como atípicos en el contexto de este conjunto de datos. Una forma de considerar esta información, podría ser clasificar a los clientes en dos categorías, aquellos con mensajes y aquellos sin mensajes, pero obtenemos exactamente la misma proporción de datos que para la variable categórica voice.mail.plan (0:2411 - 1: 922). Por lo tanto, como sería una variable exactamente repetida, la eliminaremos del análisis.

4. **Preparación del Dataset**:

Se transformó la variable objetivo "churn" asignando el valor 1 a "True" y 0 a "False". Posteriormente, el conjunto de datos se dividió en un set de entrenamiento (75% del total) y un set de prueba (25% restante), asegurando que ambas categorías de la variable "churn" fueran proporcionales entre ambos conjuntos de datos (Figura 6).



	y	n	porcentaje
	<fct>	<int>	<dbl>
1	0	2135	85.4
2	1	365	14.6

	y	n	porcentaje
	<fct>	<int>	<dbl>
1	0	715	85.8
2	1	118	14.2

Figura 6. División de los datos en set de entrenamiento (derecha) y set de prueba (izquierda).

No se emplearon técnicas de balanceo de clases para la variable “churn” en esta etapa, ya que se decidió evaluar primero el rendimiento del modelo con los datos originales. Dependiendo de los resultados obtenidos, se considerará aplicar técnicas de balanceo en etapas posteriores.

5. **Modelo Predictivo**:

Se realizó una breve revisión bibliográfica sobre los modelos implementados para predecir la fuga de clientes, encontrando que los árboles de decisión y los modelos de random forest destacan por su desempeño superior en esta tarea. Los árboles de decisión, en particular, ofrecen buenos resultados y se caracterizan por su simplicidad en la interpretación, lo cual es valioso para identificar patrones de fuga. Además, ambos modelos muestran menor sensibilidad a la presencia de datos atípicos, lo que contribuye a su robustez en entornos de datos variados (Manzoor, 2024; Kumar, 2021; Contreras et al., 2017; Pérez, 2014).

En este estudio, se entrenaron dos modelos: Random Forest y Árboles de Decisión. Este último mostró el mejor rendimiento, por lo cual se presentan sus resultados a continuación (ambos desarrollos se pueden encontrar en el script R anexo a este informe).

Se utilizó la librería rpart para entrenar un árbol de decisión, fijando el parámetro de control (cp) en 0.001. Este ajuste permite un mayor número de divisiones en el árbol, lo que puede ayudar a capturar patrones más complejos en los datos. Esta configuración se eligió para explorar la relación entre las variables y la fuga de clientes de manera más detallada.

El árbol de decisión generado tiene una profundidad de 7 niveles y cuenta con 19 nodos terminales, cada uno representando una posible clasificación final para el cliente. En los niveles superiores del árbol predominan las variables total.day.minutes, customer.service.calls y voice.mail.plan, lo cual indica su alta relevancia en la predicción de la fuga de clientes (Figura 7).

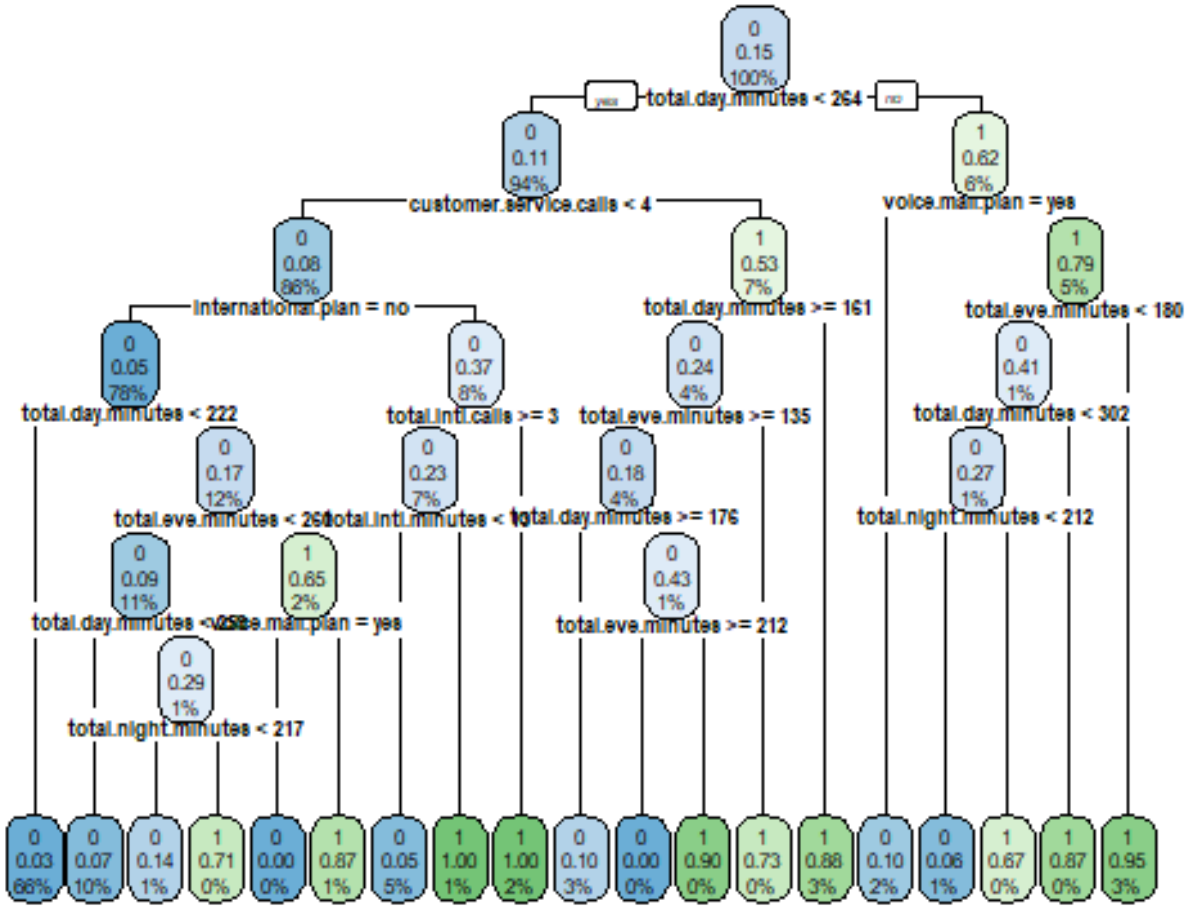


Figura 7. Visualización del árbol de decisión resultante para el primer modelo.

Las variables más importantes en el modelo corresponden al total de minutos diarios y la cantidad del costo diario, junto con la cantidad de llamadas a la mesa de ayuda que ha realizado el cliente. Las variables menos relevantes a la hora de predecir la fuga son el total de llamadas diarias, el código de área y la cantidad de llamadas en la tarde (Figura 8).

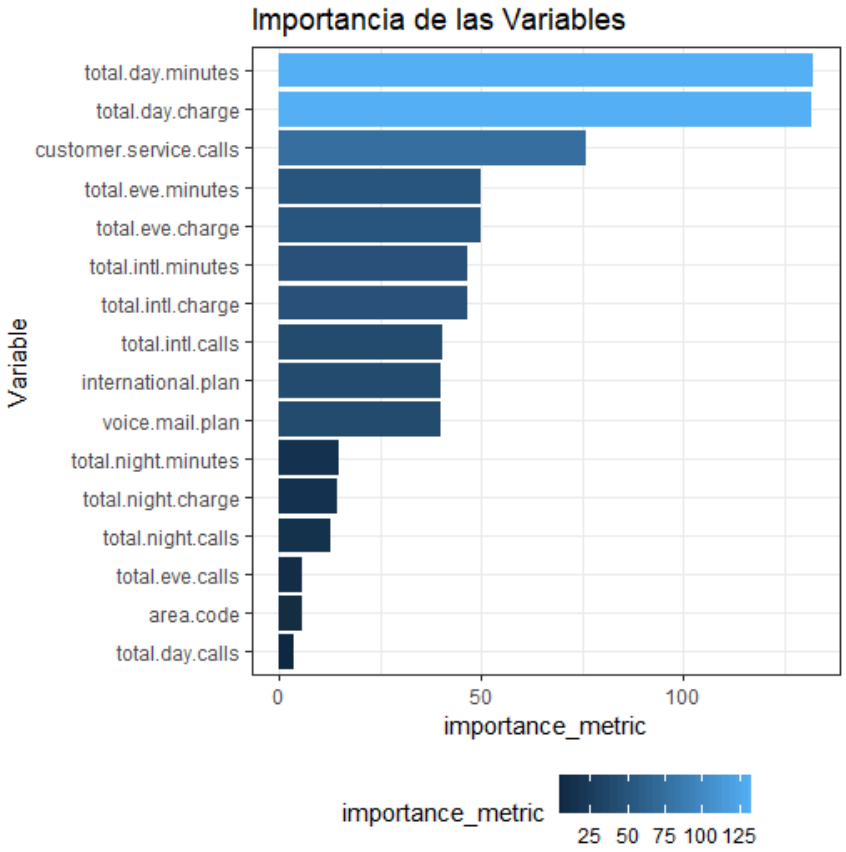


Figura 8. Ranking de la importancia de las variables del árbol de decisión.

6. Evaluación del Modelo:

El modelo fue capaz de clasificar los datos y las métricas de desempeño se muestran en la Figura 10. Estas fueron calculadas enfocándose en la clase 1, correspondiente a los clientes que se fugaron.

La matriz de confusión indica que el modelo es eficaz al identificar la mayoría de los clientes que no se fugarán, con un alto número de verdaderos negativos (699). Sin embargo, se debe prestar atención a los falsos negativos (28), ya que representan clientes que podrían haberse salvado con estrategias adecuadas de retención (Figura 10).

Referencia / Predicción	0	1
0	699	28
1	16	90

- Accuracy: 0.947
- Balanced Accuracy: 0.870
- Recall (churn = 1): 0.762
- Precision (churn = 1): 0.849
- F1 Score (churn = 1): 0.803
- AUC: 0.913

Figura 10. Matriz de confusión y principales métricas del modelo v1.

El accuracy del 94.7% sugiere que el modelo clasifica correctamente la mayoría de los casos. La precisión del 84.9% indica que, de todas las instancias clasificadas como clientes en fuga, el 84.9% son verdaderos positivos. Esto sugiere que el modelo tiene un buen control sobre las predicciones de fuga, aunque aún existe margen de mejora. Por otro lado, el recall del 76.3% refleja la proporción de verdaderos positivos identificados correctamente, lo que significa que hay un 23.7% de falsos negativos. Este aspecto es crítico en el contexto de la retención de clientes, ya que implica que el modelo puede no estar capturando todos los clientes que realmente se fugarán.

Finalmente, el AUC del 91.3% indica que el modelo posee una alta capacidad para distinguir entre los clientes que se fugarán y los que no, lo que es un resultado muy positivo en términos de rendimiento (Figura 10 y 11).

Estas métricas en conjunto sugieren que, aunque el modelo es eficaz, es esencial enfocarse en mejorar el recall para minimizar la cantidad de falsos negativos, lo que podría tener un impacto significativo en la retención de clientes.

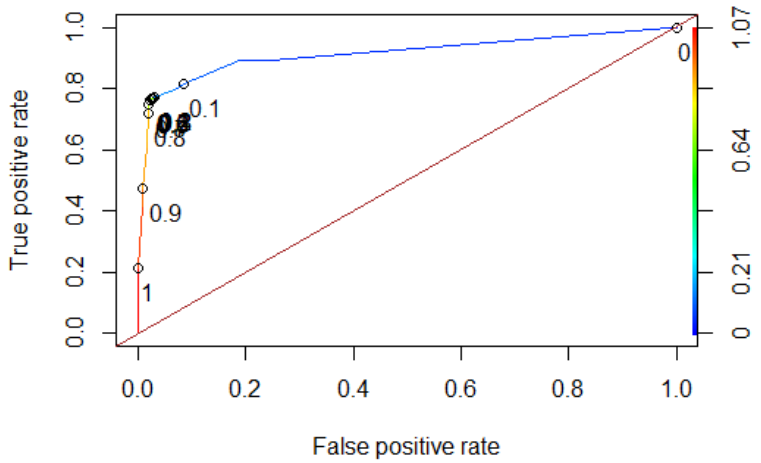


Figura 11. Curva ROC correspondiente al primer modelo de árboles de decisión.

7. Iteración y Ajuste Final:

Con el fin de optimizar el árbol de decisión, se realizó el ajuste del parámetro de control (cp), encontrando que el valor óptimo que minimizaba el error era 0.005479452. Aplicando este valor, se entrenó una segunda versión del modelo, que resultó en una estructura un poco más compacta y eficiente. Este árbol podado presenta una profundidad reducida de 6 niveles y cuenta con 22 nodos terminales (Figura 12), lo que implica una clasificación más ágil y posiblemente un menor riesgo de sobreajuste, manteniendo una alta capacidad predictiva.

Esta reducción en las divisiones contribuyó a mejorar levemente el accuracy, la precisión y el F1 score. Sin embargo, el recall y el AUC mostraron una ligera disminución (figura 13). Este comportamiento sugiere que, aunque el modelo podado es un poco más sencillo y mejora en ciertos aspectos de precisión, puede estar capturando menos casos positivos verdaderos, lo cual es un aspecto importante a considerar en el contexto de la predicción de la fuga de clientes.

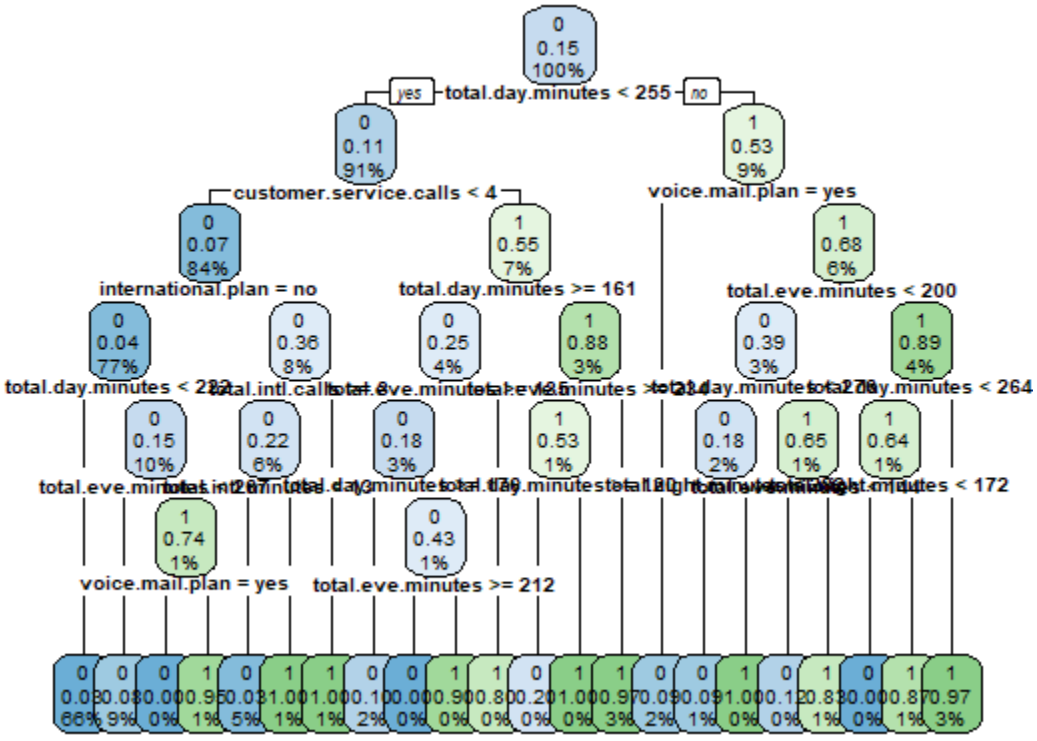


Figura 12. Visualización del árbol de decisión resultante para la v2 del modelo (poda).

Referencia / Predicción	0	1
0	709	33
1	6	85

- Accuracy: 0.953
- Balanced Accuracy: 0.855
- Recall (churn = 1): 0.720
- Precision (churn = 1): 0.934
- F1 Score (churn = 1): 0.813
- AUC: 0.906

Figura 13. Matriz de confusión y principales métricas del modelo v2 (poda).

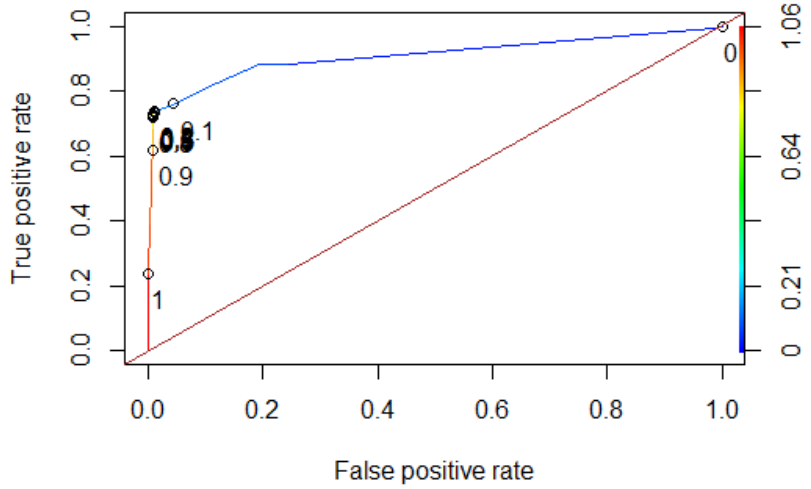


Figura 14. Curva ROC correspondiente a la segunda versión del modelo de árboles de decisión (poda).

8. Conclusiones

El análisis de la fuga de clientes ha demostrado la eficacia del modelo de árboles de decisión en la predicción de este fenómeno. Los resultados obtenidos revelan que el modelo podado del árbol de decisión no sólo mejora ligeramente las métricas de rendimiento, como el accuracy, la precisión y el F1 score, sino que también ofrece una estructura algo más simple y fácil de interpretar. Esta característica es especialmente relevante en un contexto empresarial, donde la comprensión de los factores que influyen en la fuga de clientes puede guiar estrategias de retención más efectivas.

No obstante, es fundamental considerar la ligera disminución en el recall y el AUC, ya que estas métricas son cruciales para capturar adecuadamente a los clientes en riesgo de fuga. Dado que la prioridad en este contexto es minimizar la fuga de clientes, el modelo sin poda, que mantiene un mayor recall y AUC, se presenta como una opción más adecuada. A pesar de que el modelo podado pueda parecer más atractivo por su simplicidad y sus mejoras en otras métricas, la pérdida de capacidad para identificar correctamente a los clientes en riesgo de fuga puede tener consecuencias más graves para el negocio. Por lo tanto, es razonable optar por el primer modelo sin poda, dado que ofrece un equilibrio más favorable entre la precisión de las predicciones y la necesidad de capturar a los clientes en riesgo.

Por otro lado, la importancia asignada a variables como el total de minutos diarios y el número de llamadas a la mesa de ayuda resalta que la calidad de la atención al cliente y el uso del servicio son factores determinantes en la decisión de los usuarios de continuar o no con la empresa. Esta información puede ser clave para implementar acciones que fortalezcan la fidelización de los clientes y optimicen los recursos de atención.

Referencias

Contreras, E., Ferreira, F., Valle, M. (2017). Diseño de un modelo predictivo de fuga de clientes utilizando árboles de decisión. Revista Ingeniería Industrial - Universidad del Bío Bío. http://revistas.ubiobio.cl/index.php/RI/article/view/3055

Kumar, S. (2021). Opportunities of Machine Learning on Telecom Sector: A Case Study at BSNL. International Journal of Research in Engineering and Science (IJRES), 9, 37-44.

Manzoor, A., Qureshi, M. A., Kidney, E., & Longo, L. (2024). *A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners*. [Nombre de la revista o conferencia, si corresponde].

Pérez Villanueva, P. (2014). Modelo de predicción de fuga de clientes de telefonía móvil post pago. Disponible en http://repositorio.uchile.cl/handle/2250/115942