Extracción de Conocimiento en Bases de Datos

Prevención de cancelación de contratos en compañías de telecomunicaciones

Daute Rodríguez Rodríguez

17 de junio de 2020

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Motivación y objetivos

Existe un amplio conjunto de compañías de diversa índole que hacen del análisis de la pérdida de clientes (*customer attrition* o *customer churn*) una muy importante métrica de negocio. Esto se debe al hecho de que para dichas compañías el costo de adquirir nuevos clientes es mucho mayor al de conseguir que los que ya tienen no cancelen los servicios contratados [4].

Las compañías telefónicas, los proveedores de Internet y las compañías de televisión de pago son algunos ejemplos de tipo de compañías para las cuales la retención de clientes supone un gran reto. Esto ocurre debido a que los clientes de este tipo de empresas son el principal activo de las mismas, por ello, es común que estas compañías dispongan de un servicio de atención al cliente que intente **retener o recuperar a aquellos clientes que han dado indicios de que van a cancelar los servicios contratados**.

Las empresas suelen hacer la siguiente distinción a la hora de estudiar la pérdida de clientes:

- **Baja voluntaria**: Se produce cuando el cliente decide cambiar a otra compañía o proveedor de servicios.
- **Baja involuntaria**: Se produce debido a ciertas circunstancias externas relacionadas con el cliente como la muerte o la reubicación a otro país.

Por norma general, en el momento de realizar los análisis de pérdidas de clientes se excluyen las bajas involuntarias. Las compañías centran sus estudios en las bajas voluntarias con el propósito de entender la causa de las mismas para poder actuar en consecuencia de manera preventiva. Analizar e identificar los factores que han influido en la decisión de los clientes de cancelar sus contratos con la compañía es de gran interés pues permitiría determinar la propensión o el riesgo de que otros clientes vayan a abandonar los servicios contratados.

Una vez se hayan identificado los clientes que puede que abandonen la compañía, será posible centrar los esfuerzos en retenerlos con programas de marketing específicos que incluyan contramedidas a las carencias o factores negativos detectados durante la fase de análisis.

El objetivo de este proyecto consiste en, a partir del estudio del fenómeno de pérdida de clientes, obtener un modelo de clasificación fiable capaz de anticipar si un cliente va a cancelar los servicios que tiene contratados. Esta clasificación o **predicción** se realizará atendiendo a las características propias del cliente y del contrato que el cliente mantiene con la compañía, para ello, se cuenta con un conjunto de datos [1] históricos etiquetados de una compañía de telecomunicaciones.

Esta memoria recoge los resultados obtenidos durante el desarrollo del proyecto. En concreto, el capítulo 2 describe el conjunto de datos utilizado así cómo las particularidades de cada atributo del mismo. En el capítulo 3 se muestran las tareas de limpieza, transformación y adecuación realizadas sobre el conjunto de datos con el propósito de poder estudiarlos y aplicar sobre ellos los algoritmos y técnicas oportunas. El capítulo 4 presenta el análisis llevado a cabo en primera instancia sobre los datos y una serie de observaciones relevantes que guiaron el proceso de construcción de modelos de clasificación. En el capítulo 5 se exponen los modelos de clasificación obtenidos así como los algoritmos utilizados para construirlos durante la búsqueda del mejor modelo que lograse el objetivo del proyecto. El capítulo 6 expone la comparativa de los modelos obtenidos realizada atendiendo a las distintas métricas que evalúan el rendimiento y funcionamiento de los mismos y por último, en el capítulo 7 se muestran las conclusiones alcanzadas.

Capítulo 2

Descripción de los datos

En este capítulo se presentan la naturaleza y características de los datos adquiridos para el análisis del fenómeno de pérdida de clientes. Como ya se ha mencionado con anterioridad, los datos de los que se dispone se corresponden con observaciones pertenecientes a una compañía de telecomunicaciones [1], cada observación incluye información y características de un cliente y del contrato que éste mantiene con la compañía.

El conjunto de datos se obtuvo de la plataforma *Kaggle* [6], cuenta con un total de 7043 observaciones de 21 atributos cada una. Dado que cada observación se corresponde con un cliente, cada una tiene una etiqueta que especifica si el cliente canceló su contrato con la compañía durante el último mes. El atributo *Churn* es el que determina la categoría o clase a la que pertenece una observación. Por tanto, el objetivo consiste en construir un modelo capaz de predecir o clasificar satisfactoriamente el valor del atributo *Churn* a partir de los demás atributos.

A continuación se describen los atributos del conjunto de datos y los posibles valores que pueden tomar. Han sido agrupados atendiendo al tipo de información que almacenan. El impacto que tienen los posibles valores de cada uno en la clasificación de las observaciones se expone en el capítulo 4 de la memoria.

2.1. Información sobre el cliente

Los siguientes atributos albergan información sobre un cliente de la compañía:

- *customerID*: Identificador único del cliente. Se codifica como una cadena de texto.
- *gender*: Indica el género del cliente. Posibles valores: {*Female, Male*}
- *SeniorCitizen*: Indica si el cliente está jubilado. Posibles valores: {*0, 1*}
- *Partner*: Indica si el cliente tiene pareja. Posibles valores: {*Yes, No*}
- *Dependents*: Indica si el cliente tiene personas a su cargo. Posibles valores: {*Yes, No*}
- *tenure*: Indica el número de meses que el cliente ha permanecido con la compañía. Toma el valor de un número entero positivo.
- *Churn*: Indica si el cliente canceló el contrato durante el último mes. Posibles valores: {*Yes, No*}

2.2. Información sobre el contrato

Los atributos que se pueden ver a continuación presentan características e información del contrato que mantienen cliente y compañía:

- *Contract*: Indica el tipo de contrato atendiendo a la temporalidad. Posibles valores: {*Month-to-month, One year, Two year*}
- *PaperlessBilling*: Indica si el cliente tiene facturación electrónica. Posibles valores: {*Yes, No*}
- *PaymentMethod*: Indica el método de pago. Posibles valores: {*Bank transfer (automatic), Credit Card (automatic), Electronic check, Mailed check*}
- *MonthlyCharges*: Indica la cuantía mensual del contrato. Toma el valor de un número real positivo.
- *TotalCharges*: Indica la cuantía total pagada durante el contrato. Toma el valor de un número real positivo.

2.3. Información sobre los servicios contratados

A continuación se muestran los atributos que especifican los servicios contratados por el cliente:

- *PhoneService*: Indica si el contrato incluye servicio telefónico. Posibles valores: {*Yes, No*}
- *MultipleLines*: Indica si el servicio telefónico incluye múltiples líneas de teléfono, depende del atributo *PhoneService*. Posibles valores: {*Yes, No, No phone service*}
- *InternetService*: Indica si el contrato incluye servicio de conexión a Internet. Posibles valores: {*No, DSL, Fiber optic*}
- *OnlineSecurity*: Indica si el contrato incluye servicio de seguridad online, depende del atributo *InternetSecurity*. Posibles valores: {*Yes, No, No Internet service*}
- *OnlineBackup*: Indica si el contrato incluye servicio de copias de seguridad, depende del atributo *InternetSecurity*. Posibles valores: {*Yes, No, No Internet service*}
- *DeviceProtection*: Indica si el contrato incluye el servicio de protección de dispositivos, depende del atributo *InternetSecurity*. Posibles valores: {*Yes, No, No Internet service*}
- *TechSupport*: Indica si el contrato incluye el servicio de soporte técnico, depende del atributo *InternetSecurity*. Posibles valores: {*Yes, No, No Internet service*}
- *StreamingTV*: Indica si el contrato incluye servicio de televisión en streaming, depende del atributo *InternetSecurity*. Posibles valores: {*Yes, No, No Internet service*}
- *StreamingMovies*: Indica si el contrato incluye servicio de películas en streaming, depende del atributo *InternetSecurity*. Posibles valores: {*Yes, No, No Internet service*}

Capítulo 3

Preprocesamiento de los datos

Antes de estudiar los datos y aplicar sobre ellos las técnicas de visualización y los algoritmos de construcción de modelos fue necesario llevar a cabo una serie de tareas de limpieza, transformación y adecuación. Cabe destacar que algunas de las transformaciones que a continuación se exponen se realizaron únicamente con el propósito de obtener mejores resultados a la hora de visualizar los datos. También es importante mencionar que a pesar de aplicar las transformaciones pertinentes, en ningún momento se perdieron o sobrescribieron los datos originales.

3.1. Cambio de nombre de atributos

Dado que la gran mayoría de nombres de atributos comienzan en mayúscula se creyó conveniente cambiar el nombre de los atributos *customerID*, *gender* y *tenure* por *CustomerID*, *Gender* y *Tenure* respectivamente.

3.2. Conversión de tipos de datos

Teniendo en cuenta la unicidad del atributo *CustomerID* se vio necesario cambiar el tipo de dato de *factor* a *character*. *SeniorCitizen* es el segundo y último atributo cuyo tipo de dato fue cambiado, en el conjunto de datos original se codifica como un entero pero se consideró oportuno cambiarlo a *factor* pues el resto de variables categóricas se codifican con este tipo de datos.

3.3. Cambio de valores de atributos categóricos

En esta fase del preprocesamiento de los datos es posible diferenciar dos tipos de cambios:

- Cambio de nombre de categorías de atributos categóricos
- Alteración de la cantidad de categorías de atributos categóricos

Mientras que los cambios pertenecientes al primer tipo se realizaron con el propósito de conseguir que las figuras generadas para visualizar los datos fueran más adecuadas, los cambios pertenecientes al segundo tipo se llevaron a cabo porque se creyó conveniente de cara a aplicar los algoritmos de construcción de modelos. A continuación se listan los cambios realizados correspondientes al primer tipo:

- *Churn*: {*Yes → Churn, No → Not churn*}
- *SeniorCitizen*: {*1 → Yes, 0 → No*}
- *PaymentMethod*: {*Bank transfer (automatic) → Bank transfer(A), Credit card (automatic) → Credit card(A), Electronic check → Electronic check, Mailed check → Mailed check*}
- *Contract*: {*Month-to-month → Monthly, One year → One year, Two year → Two year*}

Los cambios efectuados para alterar la cantidad de categorías de un atributo categórico son los siguientes:

- Reducción de 3 categorías a 2 en el atributo *MultipleLines*. La categoría *No phone service* equivale a la categoría *No*: {*Yes → Yes, No → No, No phone service → No*}
- Reducción de 3 categorías a 2 en el atributo *OnlineSecurity*. La categoría *No Internet service* equivale a la categoría *No*: {*Yes → Yes, No → No, No Internet service → No*}
- Reducción de 3 categorías a 2 en el atributo *OnlineBackup*. La categoría *No Internet service* equivale a la categoría *No*: {*Yes → Yes, No → No, No Internet service → No*}
- Reducción de 3 categorías a 2 en el atributo *DeviceProtection*. La categoría *No Internet service* equivale a la categoría *No*: {*Yes → Yes, No → No, No Internet service → No*}
- Reducción de 3 categorías a 2 en el atributo *TechSupport*. La categoría *No Internet service* equivale a la categoría *No*: {*Yes → Yes, No → No, No Internet service → No*}
- Reducción de 3 categorías a 2 en el atributo *StreamingTV*. La categoría *No Internet service* equivale a la categoría *No*: {*Yes → Yes, No → No, No Internet service → No*}
- Reducción de 3 categorías a 2 en el atributo *StreamingMovies*. La categoría *No Internet service* equivale a la categoría *No*: {*Yes → Yes, No → No, No Internet service → No*}

3.4. Tratamiento de valores nulos

A pesar del gran número de observaciones y atributos del conjunto de datos seleccionado sólo existen 11 valores nulos. Todos ellos se corresponden con valores del atributo *TotalCharges*. Aunque se trata de una cantidad menor (equivale al 0,0015 % del total de observaciones) se ha optado por imputar los valores faltantes en vez de eliminar las observaciones al completo. Para la imputación de los valores nulos se decidió hacer uso de la mediana del atributo.

3.5. Categorización de atributos

La categorización de atributos se estimó oportuna a fin de poder hacer uso de los mismos en los métodos de construcción de modelos que sólo soportan atributos categóricos y con el propósito de poder estudiar e interpretar de manera más simple los atributos [5]. A continuación se muestran aquellos atributos que se categorizaron y el método seleccionado:

- *Tenure*: Dado que la información que alberga este atributo es la cantidad de meses que el cliente lleva con la compañía, la categorización se llevó a cabo haciendo uso de intervalos de igual anchura (12 meses, o lo que es lo mismo, por años). Se obtuvo el mínimo (0) y el máximo (72) valor del atributo y se creó un nuevo atributo denominado *CatTenure*. Los posibles valores de este nuevo atributo son: { 0-1 year, 1-2 years, 2-3 years, 3-4 years, 4-5 years, 5-6 years}
- *MonthlyCharges*: Se optó por crear 3 categorías para el nuevo atributo *CatMonthlyCharges*: { Low, Medium, High}. Al igual que con el atributo *CatTenure* los intervalos que definen la categoría tienen la misma anchura.
- *TotalCharges*: Al igual que con el atributo *MonthlyCharges*, se definieron 3 categorías para el nuevo atributo *CatTotalCharges* ({ Low, Medium, High}) utilizando intervalos de igual anchura.

3.6. Normalización de atributos

Con el propósito de evitar que los atributos numéricos con una escala mayor predominaran sobre los atributos con escalas menores [2] a la hora de aplicar los algoritmos de obtención de modelos se llevó a cabo la normalización de los mismos. Los atributos normalizados fueron *Tenure*, *MonthlyCharges* y *TotalCharges*. Para la normalización se aplicó la estrategia *z-score*.

3.7. Creación de atributos *dummy*

Algunos modelos como los de regresión logística no admiten variables categóricas no binarias, o lo que es lo mismo, requieren que las variables categóricas de más de dos clases se codifiquen como variables binarias. Los atributos *dummy* son variables que únicamente pueden tomar los valores 0 o 1 e indican la ausencia o presencia de un efecto categórico [3]. Un atributo independiente *dummy* que dada una observación toma el valor 0 conseguirá que el coeficiente asociado a dicho atributo no influya en la variable dependiente. Por el contrario, cuando el atributo *dummy* toma el valor 1, la presencia del efecto categórico influirá en el resultado atendiendo al coeficiente asociado al atributo.

A partir de una variable categórica con n clases se obtienen $n-1$ atributos *dummy*. Cada uno representa una posible clase del atributo categórico original, la clase faltante se representa mediante el valor 0 en todos los atributos *dummy*. Dentro del conjunto de datos, los atributos categóricos sustituidos por sus respectivos atributos *dummy* fueron:

- *InternetService* $\rightarrow$ {*InternetService DSL*, *InternetService Fiber optic*}
- *Contract* $\rightarrow$ {*Contract One year*, *Contract Two years*}
- *PaymentMethod* $\rightarrow$ {*PaymentMethod Credit card (A)*, *PaymentMethod Electronic check*, *PaymentMethod Mailed check*}
- *CatTenure* $\rightarrow$ {*CatTenure 1-2 years*, *CatTenure 2-3 years*, *CatTenure CatTenure 3-4 years*, *CatTenure CatTenure 4-5 years*, *CatTenure CatTenure 5-6 years*}
- *CatMonthlyCharges* $\rightarrow$ {*CatMonthlyCharges Medium*, *CatMonthlyCharges High*}

Capítulo 4

Análisis exploratorio

A lo largo de este capítulo se presentan una serie de reflexiones iniciales extraídas del análisis de la distribución de valores de los atributos frente al valor que toma el atributo *Churn*. Estas reflexiones servirán de punto de partida y de guía para la fase de construcción de modelos. El análisis de la distribución de valores de los atributos se realiza siguiendo los agrupamientos presentes en el capítulo 2:

- Atributos que almacenan información sobre el cliente.
- Atributos que almacenan información sobre el contrato.
- Atributos que almacenan información sobre los servicios contratados.

Antes de comenzar el estudio del resto de los atributos, conviene analizar la distribución de valores del propio atributo *Churn*. En la figura 4.1 es posible apreciar el porcentaje y la cantidad exacta de clientes que cancelaron su contrato durante el último mes. Atendiendo a la distribución de valores del atributo de clase se puede deducir que los datos de los que se disponen presentan un problema de clases desbalanceadas. Este hecho habrá de tenerse en cuenta a la hora de generar y evaluar los modelos de clasificación pues en caso contrario puede llegar a influir negativamente en el rendimiento de los mismos.

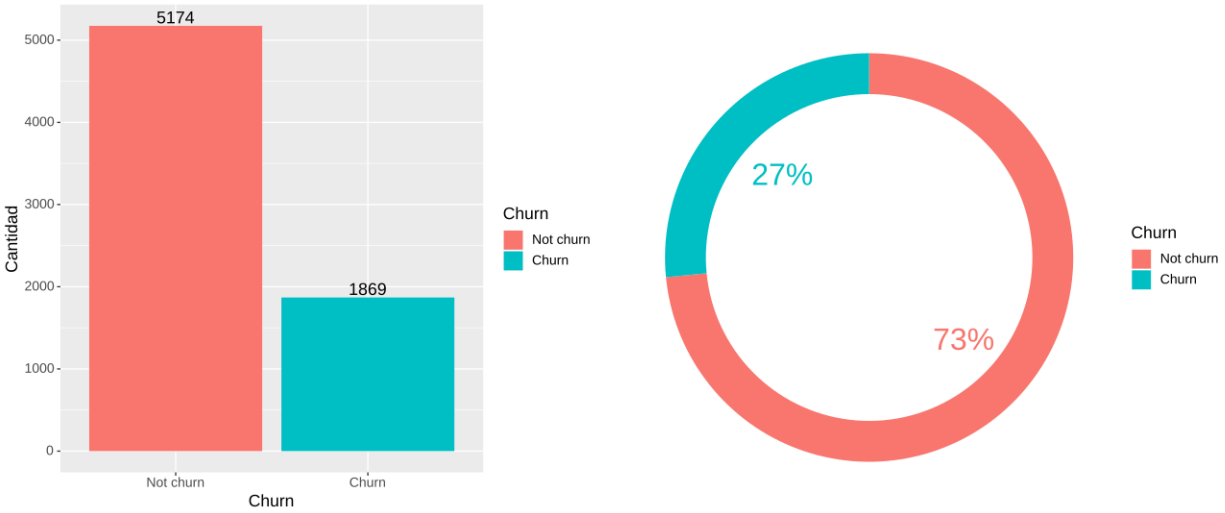

Figura 4.1: Distribución de valores del atributo *Churn*

4.1. Atributos con información sobre el cliente

A continuación se expone el análisis de aquellos atributos que almacenan información sobre el cliente. Atendiendo a los valores que se muestran en la figura 4.2 se puede concluir que el género de un cliente no influye en el posible abandono de la compañía por parte del mismo. Por el contrario, se puede apreciar una diferencia significativa a la hora de estudiar el atributo *SeniorCitizen*. Al parecer los clientes jubilados son más propensos a abandonar la compañía que los no jubilados, la relación de los porcentajes de jubilados que abandonan frente a los jubilados que permanecen con la compañía es del doble.

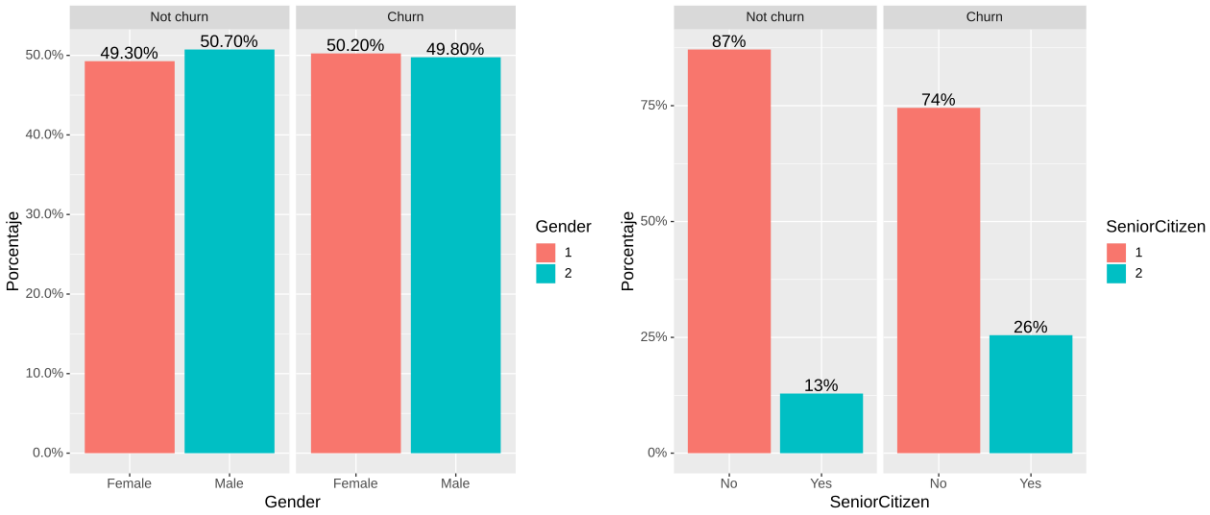

Figura 4.2: Distribución de los atributos *Gender* y *SeniorCitizen* frente a *Churn*

Entre los clientes con personas a su cargo o con pareja ocurre lo contrario que en el caso de los clientes jubilados. La figura 4.3 muestra cómo los porcentajes de cancelación de contrato en este tipo de clientes son menores que en clientes sin estos tipos de situación.

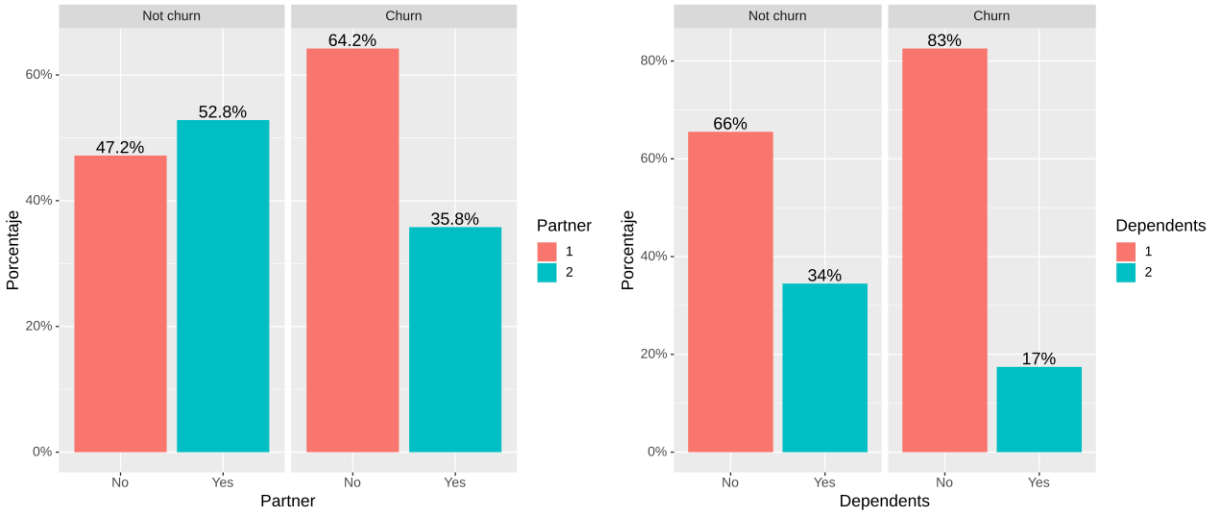

Figura 4.3: Distribución de los atributos *Partner* y *Dependents* frente a *Churn*

Por último, resta estudiar el atributo *Tenure* dentro del grupo de atributos que almacenan información sobre el cliente, en la figura 4.4 se presenta la distribución de valores que toma este atributo en el conjunto de datos. Es posible apreciar cómo los valores están repartidos de manera bastante uniforme en el caso de las observaciones correspondientes a clientes que no cancelaron su contrato (*Not churn*). Sin embargo, para las observaciones de clientes que sí que lo cancelaron (*Churn*), ocurre lo contrario, una amplia mayoría del total de valores presentes en la distribución puede encontrarse en el rango de 0 a 24 meses.

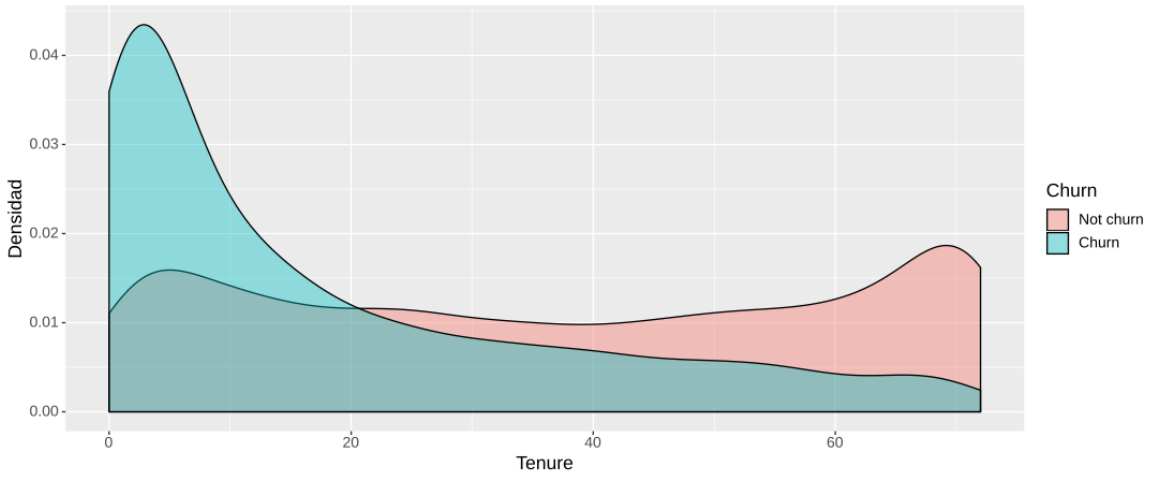

Figura 4.4: Distribución del atributo *Tenure* frente a *Churn*

Prestando atención a la figura 4.5 se corrobora la afirmación previamente realizada. La mayoría de clientes (alrededor del 70 %) que abandonó la compañía durante el último mes tan sólo llevaban con la compañía de 0 a 24 meses. Por norma general, los clientes a los que les gusta la calidad de los servicios que una compañía ofrece se mantienen con ella. Por ello es entendible que de los clientes que abandonaron la compañía un porcentaje menor se corresponda con clientes con varios años de antigüedad. A pesar de esto, resulta sorprendente que el porcentaje de clientes que llevan un año o menos con la compañía y la abandonan sea tan alto.

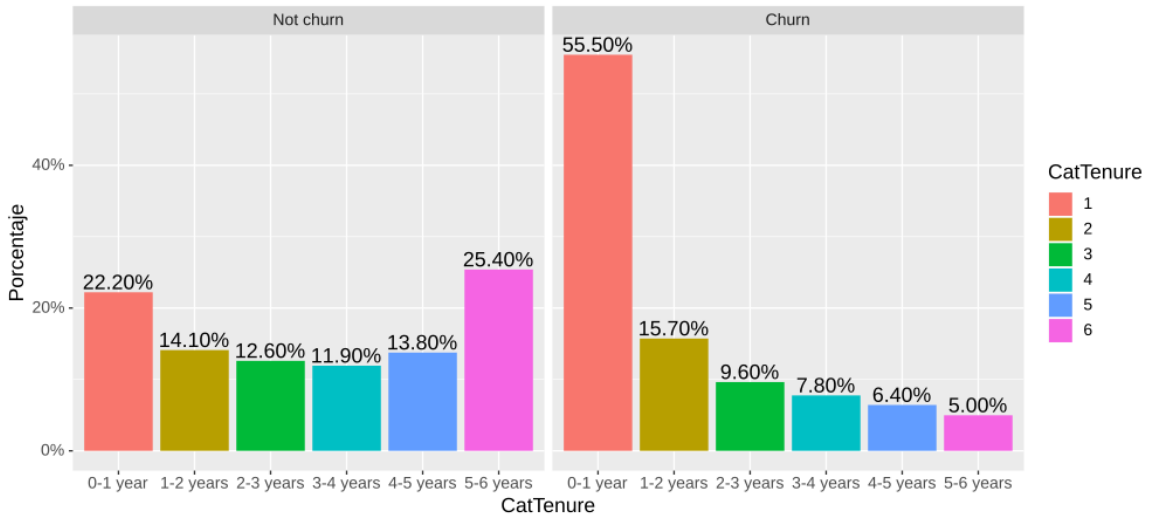

Figura 4.5: Distribución del atributo *CatTenure* frente a *Churn*

4.2. Atributos con información sobre el contrato

En esta sección se presenta el estudio de los atributos que guardan relación con el contrato que el cliente mantiene con la compañía. La figura 4.6 muestra los porcentajes de clientes que la abandonaron para cada uno de los posibles valores de los atributos *Contract* y *PaperlessBilling*. El porcentaje de clientes con contrato mensual que lo han cancelado durante el último mes es inmenso en comparación al de los clientes con este tipo de contrato que no abandonaron la compañía. Lo cierto es que para aquellos clientes que mantienen contratos mes a mes resulta menos difícil cancelar el contrato si así lo desean, los contratos de un año o dos suelen conllevar periodos de permanencia que obligan al cliente a permanecer con la compañía. Además, atendiendo a la información de la figura 4.7 podemos comprobar que la proporción de clientes con contratos mes a mes es superior a la de clientes con contratos de un año o dos.

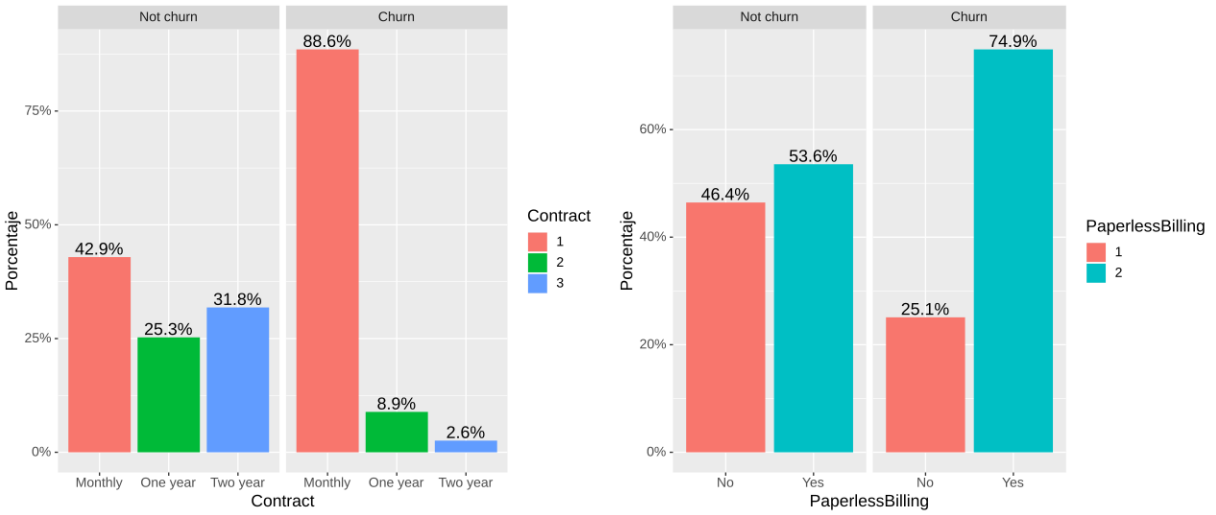

Figura 4.6: Distribución de los atributos *Contract* y *PaperlessBilling* frente a *Churn*

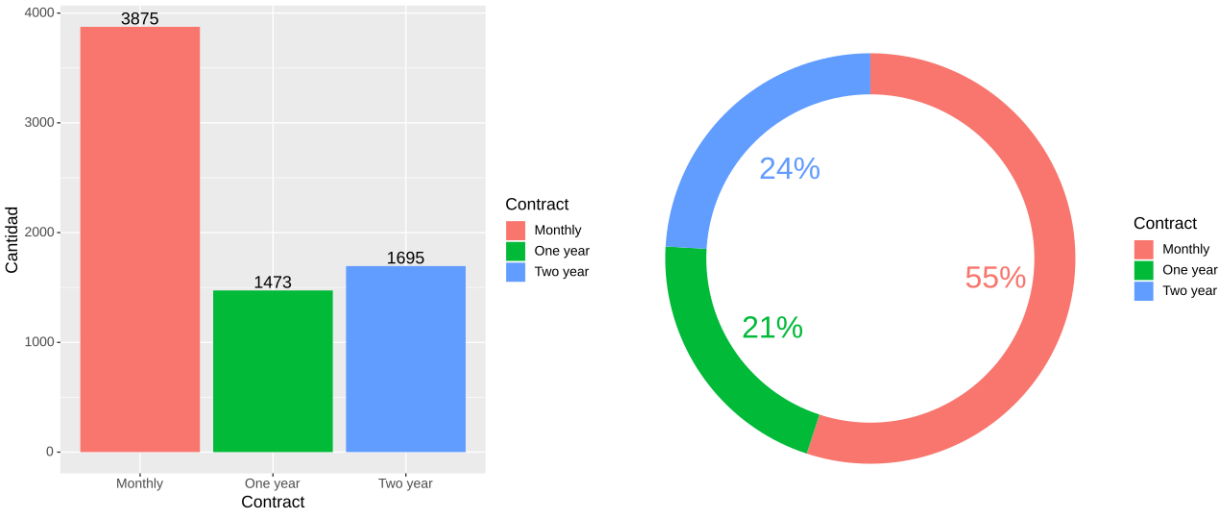

Figura 4.7: Distribución de valores del atributo *Contract*

Volviendo a la figura 4.6, resta mencionar la distribución de valores del atributo *PaperlessBilling* frente a la variable de clase. Al parecer, los clientes con facturación electrónica son más

propensos a abandonar la compañía pues 3 de cada 4 (aproximadamente) clientes que abandona hacen uso de este tipo de facturación.

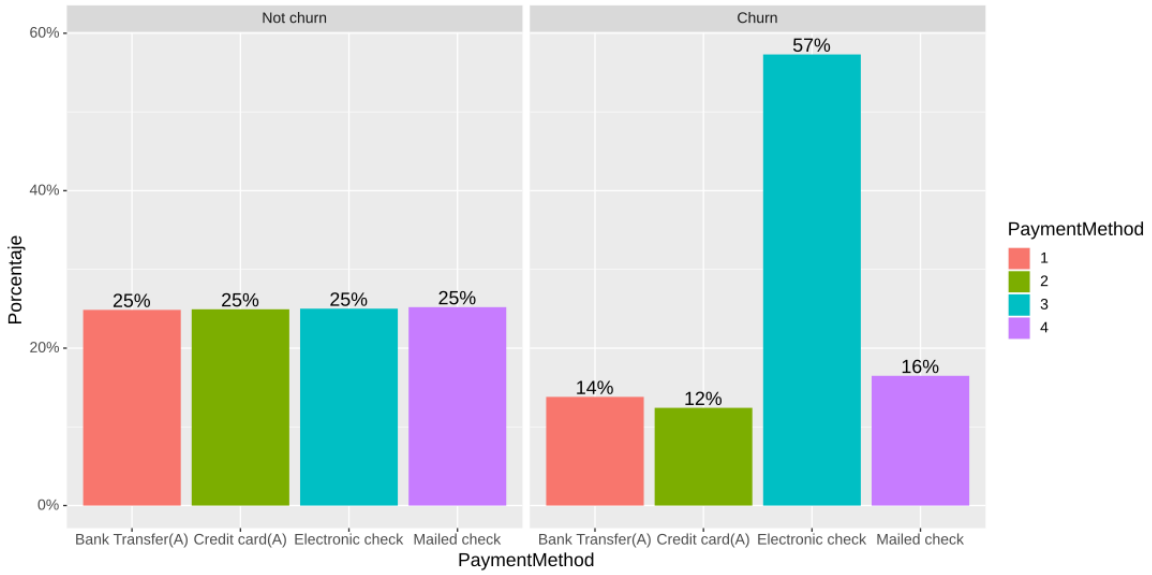

Figura 4.8: Distribución del atributo *PaymentMethod* frente a *Churn*

La figura 4.8 muestra que entre los clientes que no cancelaron su contrato se hace aproximadamente el mismo uso de los distintos métodos de pago admitidos. Sin embargo, atendiendo a los clientes que sí que abandonaron, es posible apreciar cómo un amplio número utiliza el método de pago por cheque electrónico. Al estudiar el atributo *MonthlyCharges* (figuras 4.9 y 4.10) se advierte cierta tendencia al abandono cuando la cuantía mensual supone una cantidad considerable (media-alta). Conociendo la alta competencia existente entre las compañías que proveen servicios telefónicos y de acceso a Internet, esto puede deberse a que los clientes reciben ofertas de otras compañías que igualan o mejoran los servicios que ya tienen contratados por una cuota menor.

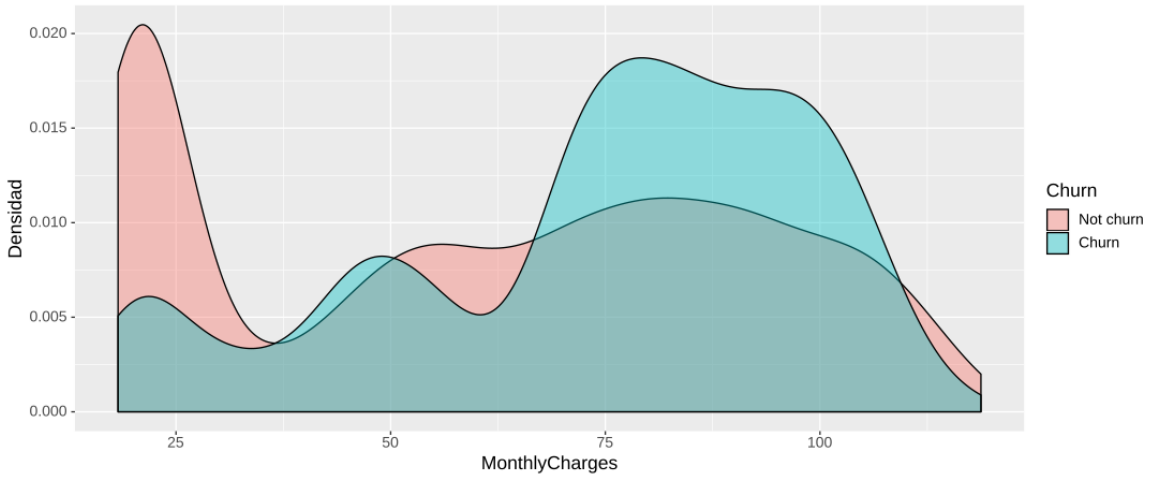

Figura 4.9: Distribución del atributo *MonthlyCharges* frente a *Churn*

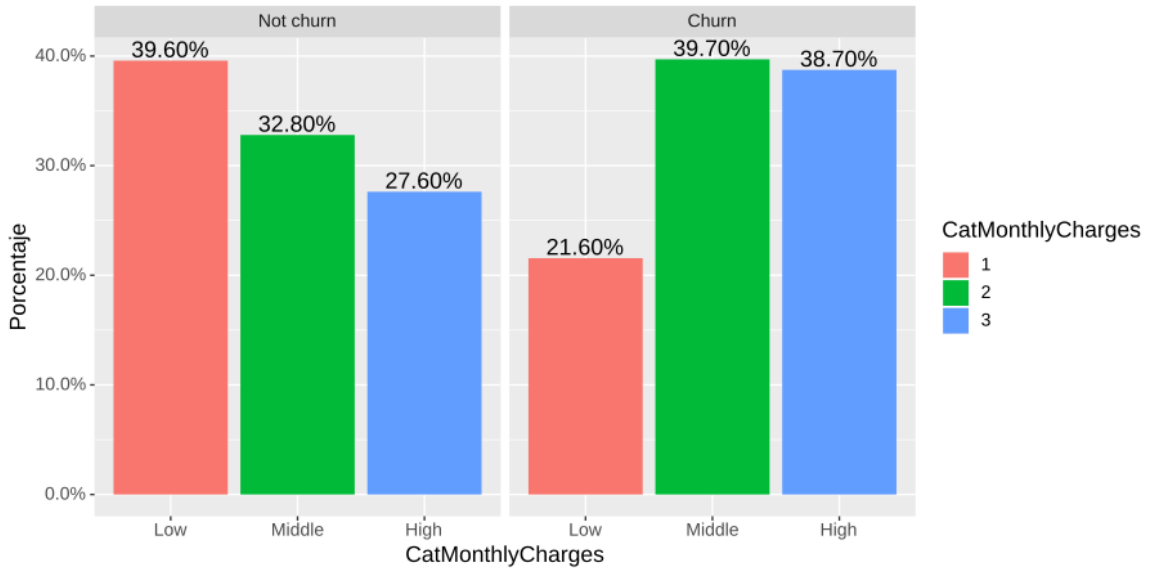

Figura 4.10: Distribución del atributo *CatMonthlyCharges* frente a *Churn*

También es posible advertir que para cuantías mensuales bajas la relación de porcentajes de clientes que permanecen y clientes que abandonan es de la mitad aproximadamente. Siguiendo con el argumento previamente expuesto, es posible considerar que para los clientes con cuantías bajas no merezca la pena realizar un cambio de compañía pues la posible mejora en el precio por los mismos servicios es baja.

En la figura 4.11 se puede ver cómo la cantidad total pagada por los clientes que abandonan la compañía suele ser menos. Esto puede deberse a que la mayoría de clientes que abandona la compañía lo hacen cuando llevan poco tiempo, tal y como se comentó anteriormente al estudiar la información que proporciona la figura 4.5.

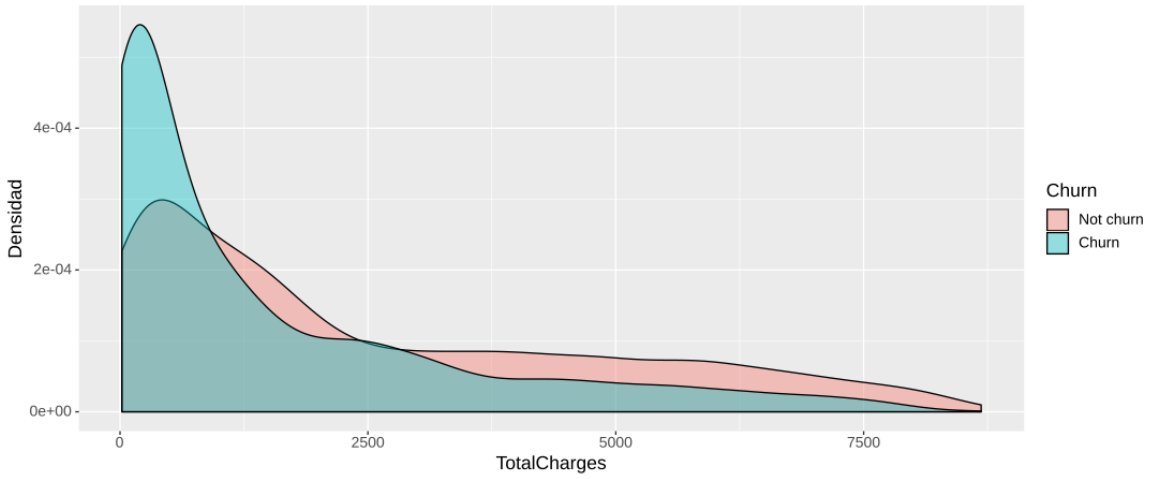

Figura 4.11: Distribución del atributo *TotalCharges* frente a *Churn*

4.3. Atributos con información sobre los servicios contratados

A lo largo de esta sección se exponen las reflexiones extraídas del análisis de la distribución de valores de los atributos con información sobre los servicios contratados frente a *Churn*. Cabe destacar que en este grupo de atributos existen dos subgrupos, cada uno se relaciona con uno de los servicios principales que la compañía ofrece: servicio de teléfono y de conexión a Internet.

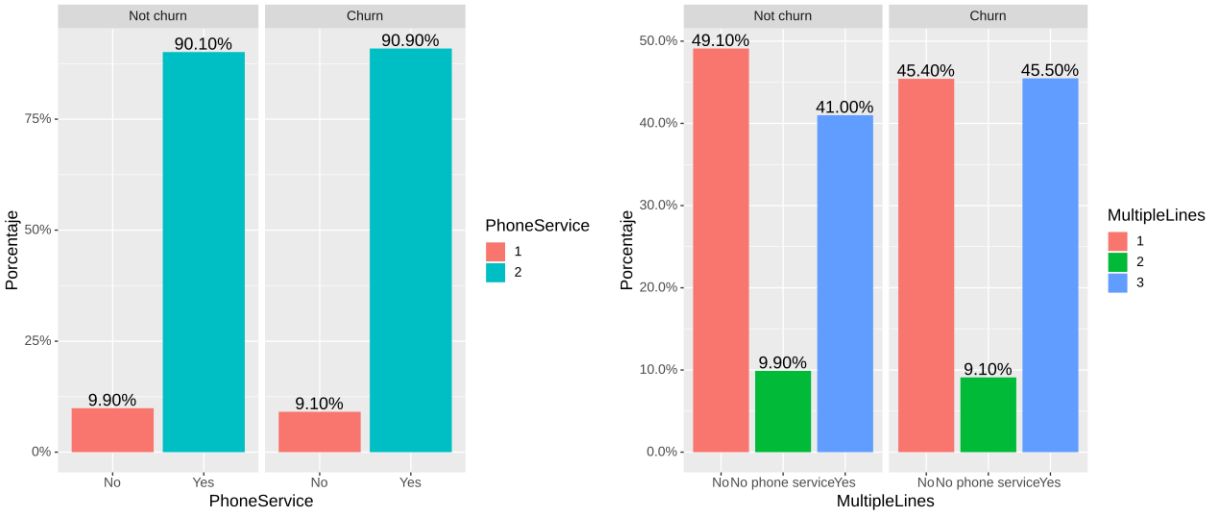

Figura 4.12: Distribución de los atributos *PhoneService* y *MultipleLines* frente a *Churn*

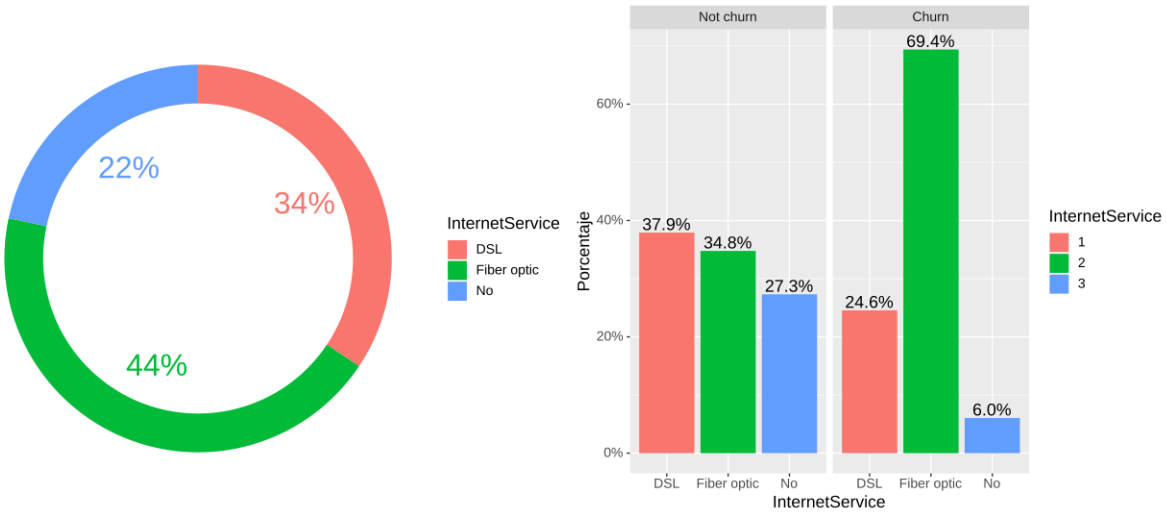

Figura 4.13: Distribución del atributo *InternetService* frente a *Churn*

La figura 4.12 muestra el porcentaje de clientes que tienen contratado el servicio de teléfono y el de múltiples líneas telefónicas. En ninguno de ambos casos parece existir relación con la tasa de abandono pues los porcentajes para cada categoría de cada atributo son muy similares estudiando el conjunto de clientes que abandonan y los que permanecen.

De la información que aportan las figuras 4.12 y 4.13 pueden deducirse las siguientes afirmaciones acerca de los servicios principales de la compañía. El servicio de teléfono parece no

afectar a la tasa de cancelación ya que los porcentajes de contratación son prácticamente los mismos entre los clientes que abandonan y los que no. Por el contrario, en el servicio de conexión a Internet sí que existen diferencias, los porcentajes asociados a cada categoría del atributo en los casos de clientes que permanecen y clientes que abandonan son muy dispares. En concreto, la diferencia más significativa y de más peso es la que se produce entre los clientes que tienen contratado el servicio de conexión a Internet por fibra óptica. Aproximadamente, 7 de cada 10 clientes que abandonan la compañía tienen contratado este tipo de conexión a Internet.

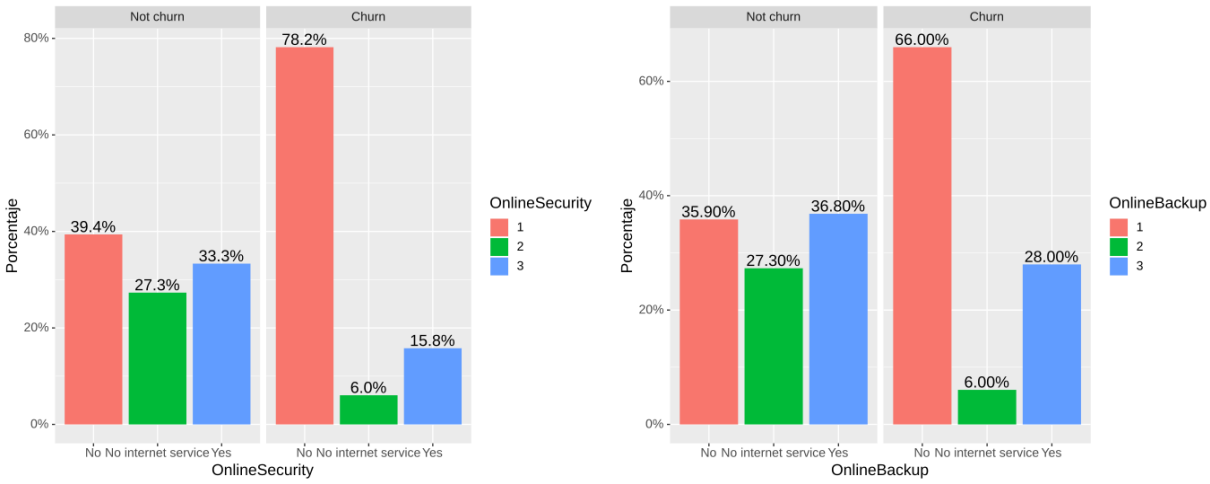

Figura 4.14: Distribución de los atributos *OnlineSecurity* y *OnlineBackup* frente a *Churn*

Los datos visibles en las figuras 4.14 y 4.15 evidencian que lo más común entre los clientes que abandonan la compañía es que no tengan contratados servicios extra aparte del servicio de conexión a Internet. Esto implica que la contratación de servicios extra hace que los clientes sean menos propensos a cancelar sus contratos.

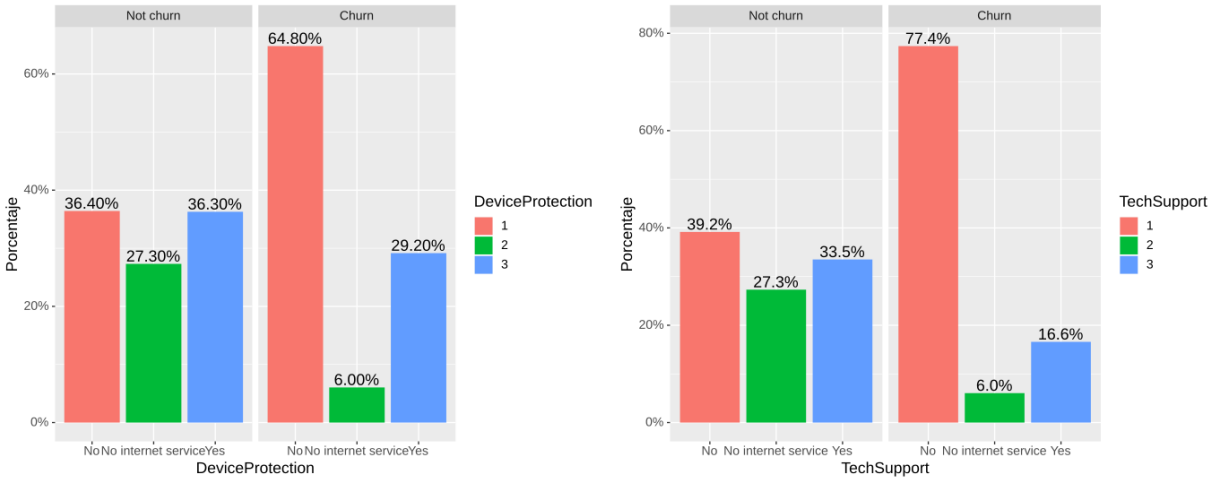

Figura 4.15: Distribución de los atributos *DeviceProtection* y *TechSupport* frente a *Churn*

Asumiendo que la relación calidad/precio de los servicios ofertados es buena y que los clientes están conformes con dichos servicios, es entendible que el porcentaje de contratación sea bajo entre los clientes que abandonan la compañía. En particular, puede llegar a llamar la atención

que en el caso de los servicios *OnlineSecurity* y *TechSupport* dicho porcentaje sea significativamente inferior al del resto de servicios relacionados con el de conexión a Internet que se ofertan. Teniendo en cuenta que alrededor de 3 clientes de cada 10 contrata estos servicios (tal y como se muestra en la figura 4.16) se pueden concluir las siguientes observaciones:

- Son servicios de excelente calidad y/o exclusivos de la compañía a los que otras compañías no pueden hacer competencia
- Son servicios que crean una dependencia importante del cliente hacia la compañía

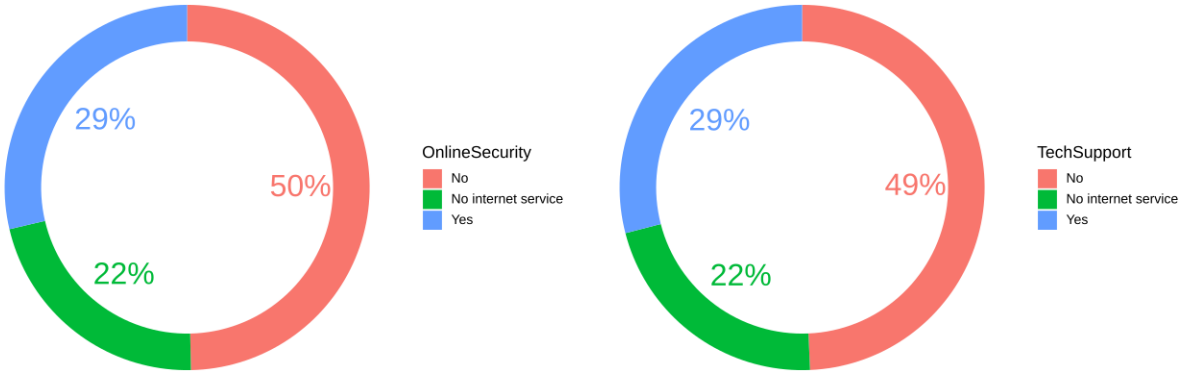

Figura 4.16: Distribución de los atributos *OnlineSecurity* y *TechSupport*

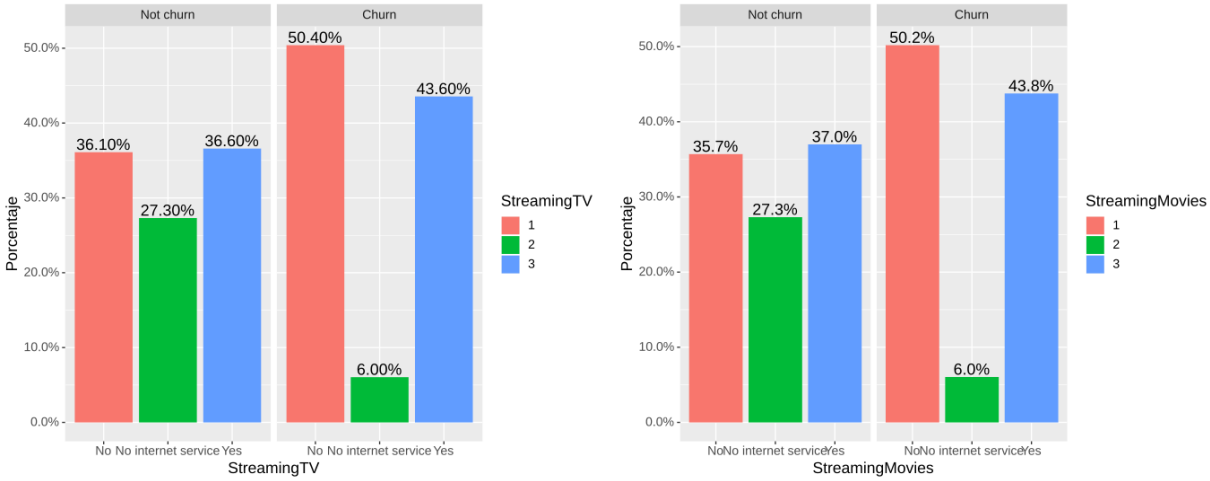

Figura 4.17: Distribución de los atributos *StreamingTV* y *StreamingMovies* frente a *Churn*

A diferencia que con el resto de servicios relacionados con el de conexión a Internet, los porcentajes de clientes que contratan los servicios *StreamingTV* y *StreamingMovies* son relativamente próximos a los de clientes que no los contratan dentro del conjunto de clientes que abandonan la compañía. A partir de esta observación se puede concluir que al parecer, el tener contratados estos servicios no influye a la hora de tomar la decisión de cancelar el contrato por parte de los clientes. Lo cierto es que este tipo de servicios disfrutan de una gran popularidad actualmente, y por ello la mayoría de compañías del sector los incluye entre sus ofertas.

Capítulo 5

Construcción de modelos

En este capítulo se muestran los modelos de clasificación construidos para determinar si un cliente abandonará o permanecerá en la compañía. Por cada modelo obtenido se presentan una serie de estadísticos que miden el rendimiento del mismo. En el capítulo 6 se compararán con el propósito de determinar qué modelo se comporta mejor ante el problema. Los estadísticos que se analizarán son los siguientes:

- *Exactitud*: Proporción de clientes correctamente clasificados
- *Tasa de verdaderos positivos o Sensibilidad*: Clientes correctamente clasificados como *Churn* entre el número total de clientes que abandonaron la compañía
- *Tasa de verdaderos negativos o Especificidad*: Clientes correctamente clasificados como *Not churn* entre el número total de clientes que no abandonaron la compañía
- *Precisión*: Clientes correctamente clasificados como *Churn* entre el número total de clientes clasificados como *Churn*
- *Índice Kappa*: Mide la concordancia existente entre las frecuencias de ocurrencia reales de cada clase (clientes que abandonan y los que no lo hacen) y las frecuencias obtenidas al usar el clasificador
- *Puntuación F1*: Media armónica de los estadísticos *True Positive rate* y *Precisión*

5.1. Regresión logística

Dada la naturaleza binaria del atributo de clase es posible construir un modelo de clasificación basado en regresión logística. La regresión logística es un modelo estadístico que en su forma básica hace uso de una función logística para modelar una variable dependiente binaria. En dicho modelo, el logaritmo de la razón de probabilidades (*log-odds*) [12] de que la variable dependiente tome el valor 1 es una combinación lineal de una o más variables independientes (predictores) binarias o continuas. Para cada variable independiente se estima un coeficiente que determinará su peso dentro del modelo. Además, sobre cada variable independiente se ejecutan tests estadísticos que determinan su nivel de significación. Cuando se le proporciona una observación como entrada al modelo, éste devuelve una probabilidad. El modelo en sí mismo modela la probabilidad dada una entrada específica pero no realiza una clasificación estadística. La regresión logística se convierte en una técnica de clasificación cuando se especifica un umbral de decisión o *cutoff* de manera que:

- Las probabilidades inferiores o iguales al umbral se clasifican como pertenecientes a la clase 0
- Las probabilidades superiores al umbral se clasifican como pertenecientes a la clase 1

Atendiendo al problema de prevención de cancelación de contratos en compañías de telecomunicaciones, cuando la probabilidad que devuelva el modelo sea mayor que el umbral especificado el cliente se clasificará como *Churn* (cliente que abandonará la compañía) y en el caso contrario como *Not churn* (cliente que permanecerá). Dado que la regresión logística admite como predictores variables binarias o continuas, para la construcción del modelo se decidió hacer uso de los atributos categóricos de dos niveles y los atributos *dummy* de los correspondientes atributos categóricos con más de dos niveles. También se hizo uso de los atributos numéricos normalizados y se excluyeron los atributos numéricos categorizados *CatTenure* y *CatMonthlyCharges*.

A modo de primera aproximación, se optó por entrenar el modelo utilizando todos los atributos disponibles. En la figura 5.1 puede observarse el gráfico de densidad de las probabilidades devueltas por el modelo teniendo en cuenta el valor real del atributo de clase. Idealmente, la intersección de las áreas que quedan bajo las curvas sería 0, de manera que se pudiese establecer un *cutoff* que separase limpiamente ambas. Al establecer el *cutoff*, aquella porción del área correspondiente a la clase *Churn* que quede a la izquierda del valor de *cutoff* establecido se corresponderá con los falsos negativos mientras que la porción del área de la clase *Not churn* que quede a la derecha se corresponderá con los falsos positivos.

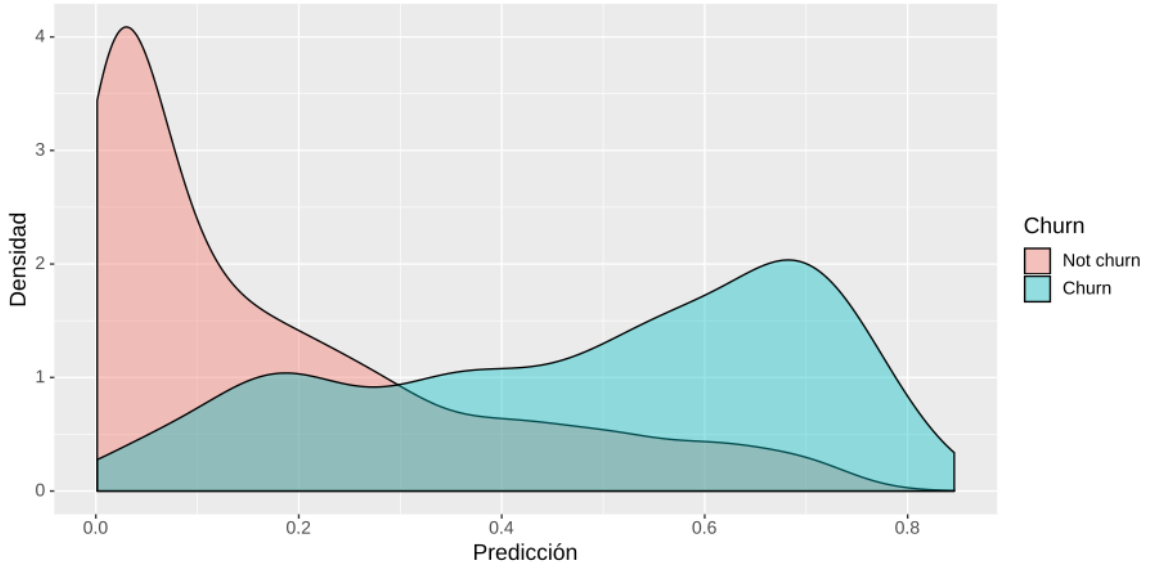

Figura 5.1: Gráfico de densidad de las predicciones realizadas por el modelo de regresión logística sobre el conjunto de validación

Para el entrenamiento de los modelos de regresión logística se optó por subdividir el número de observaciones del conjunto de datos en dos grupos teniendo en cuenta la proporción de los distintos valores del atributo de clase: entrenamiento y validación. El conjunto de entrenamiento se conformó a partir del 70 % de las observaciones disponibles y el de validación con el 30 % restante.

Tras el entrenamiento, se llevó a cabo la fase de validación y obtención de resultados. Para poder predecir nuevas observaciones es necesario especificar el *cutoff* o umbral. Lo cierto es que

la selección del *cutoff* es un problema en sí mismo pues alterará enormemente el comportamiento y la naturaleza del modelo de clasificación. Al seleccionar un *cutoff*, se está estableciendo un compromiso entre la tasa de verdaderos positivos (proporción de clientes etiquetados correctamente como *Churn*) y la tasa de verdaderos negativos (proporción de clientes etiquetados correctamente como *Not churn*), o lo que es lo mismo, se crea un balance entre los costos asociados al intento de retener clientes que no tienen pensado abandonar la compañía (falsos positivos) y los costos asociados al abandono de clientes que el modelo no detectó como clientes con intención de cancelar su contrato (falsos negativos). Por todo esto, la selección del mejor *cutoff* para el modelo se convierte en una decisión de negocio, en la que se han de tener en cuenta factores tales como el costo asociado a las diferentes acciones que desencadenará el modelo al predecir correctamente e incorrectamente. En la figura 5.2 se presenta el valor que toman los estadísticos para cada valor de *cutoff* en el rango [0.15, 0.85].

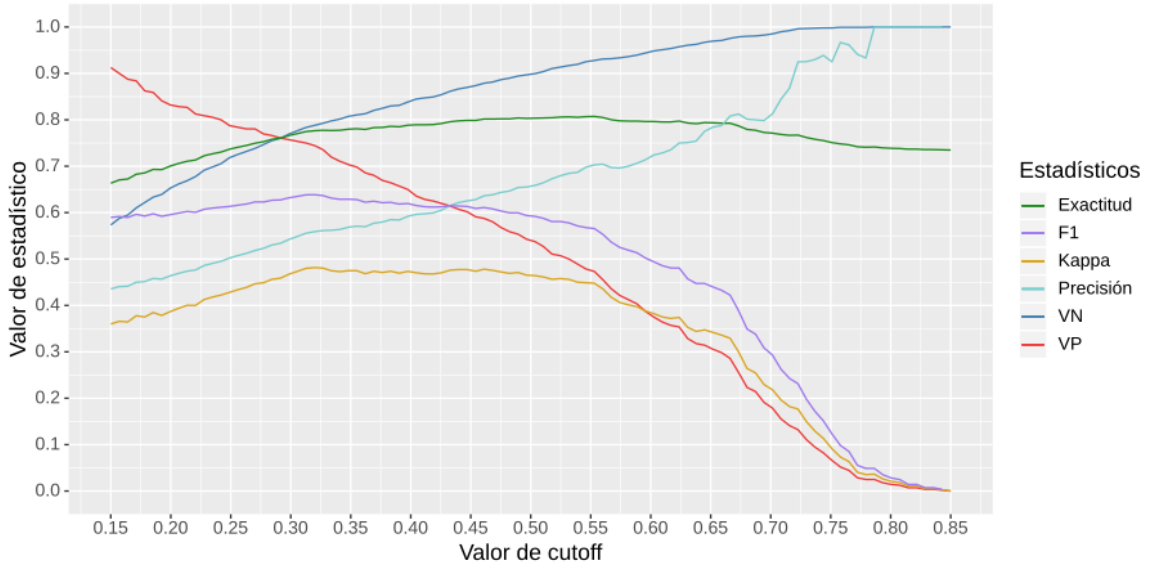

Figura 5.2: Valor que toman los estadísticos para los distintos valores de *cutoff*

Una posibilidad a la hora de seleccionar el *cutoff* para el modelo consiste en establecer el coste asociado a los falsos positivos y a los falsos negativos, de manera que sea posible encontrar el valor de *cutoff* que minimice el coste total [7]. En caso de que los costes sean distintos, se priorizará la correcta clasificación de una clase frente a la otra. Teniendo en cuenta el problema de la prevención de cancelación de contratos, debería asignarse un coste mayor a los falsos negativos, es decir, a los clientes que abandonarán la compañía y el modelo no es capaz de detectar. Sin embargo, tal y como se ha comentado con anterioridad, los costes asociados a estas situaciones variarán dependiendo de varias circunstancias como la situación de la compañía o la intensidad de los intentos de retención de clientes.

Con el propósito de ejemplificar la selección de un valor para el parámetro *cutoff* y poder estudiar el comportamiento del modelo se ha decidido asignar al coste de los falsos positivos 1 unidad y al de los falsos negativos 3 unidades. Extrapolándolo al problema, esto se traduce en que perder un cliente es 3 veces más caro que intentar retenerlo. Cabe destacar que asignando estos costes a cada tipo de error de clasificación se está consiguiendo solucionar el problema de clases desbalanceadas del conjunto de datos.

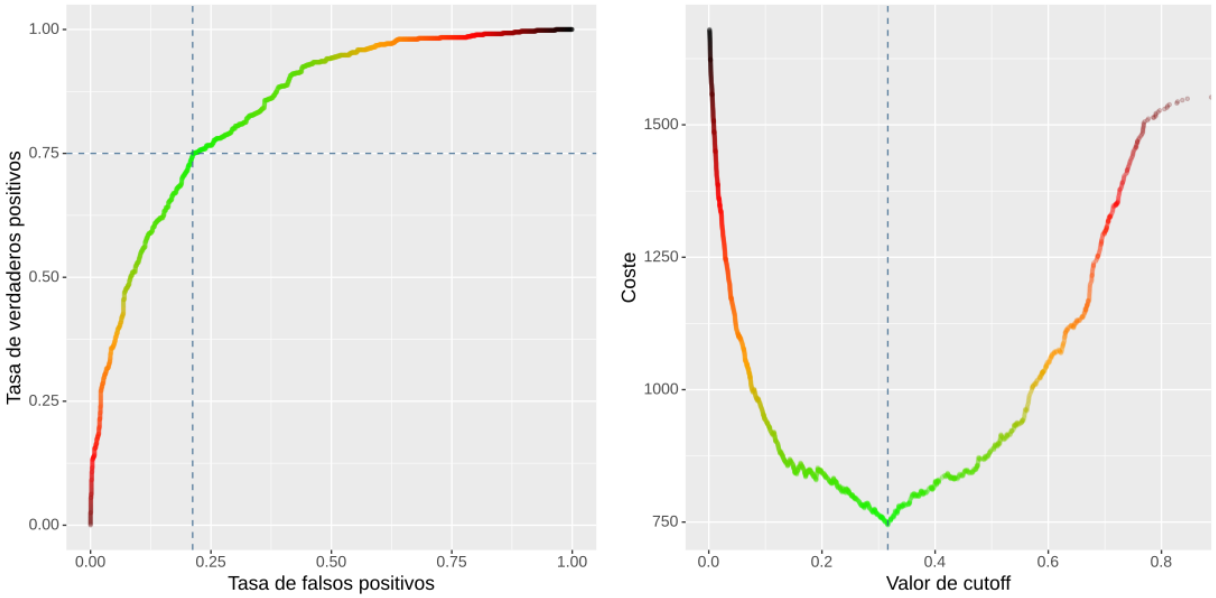

Figura 5.3: Curva ROC y coste total asociados a distintos valores de *cutoff*

Una vez asignados los valores a cada tipo de error de clasificación es posible encontrar el valor de *cutoff* que minimiza el coste total, la figura 5.3 contiene el resultado de esta búsqueda. En la gráfica de la izquierda, es posible apreciar la curva ROC del modelo para distintos valores del parámetro *cutoff*. Las líneas discontinuas de color azul reflejan los valores de las tasas de verdaderos y falsos positivos del modelo al hacer uso del valor de *cutoff* óptimo. El color de la curva representa una escala del coste total. En la gráfica de la derecha se expone la relación existentente entre los posibles valores del parámetro *cutoff* y el coste total asociado. Es posible apreciar cómo el color de la curva varía a medida que el coste total aumenta, al igual que con la gráfica de la izquierda el color de la curva representa una escala del coste total. El valor de *cutoff* que minimiza el coste total se representa por medio de una línea discontinua azul, en este caso se trata del valor 0.3164 aproximadamente. El coste mínimo asociado a este valor de *cutoff* es de 744.13 unidades aproximadamente.

	Clase real	
	Churn	*Not churn*
Churn	419	329
Not churn	141	1223

Estadístico	**Valor**
Exactitud	0.7775
Tasa de VP	0.7482
Tasa de VN	0.7880
Precisión	0.5602
Índice Kappa	0.4843
Puntuación F1	0.6407

Cuadro 5.1: Matriz de confusión y estadísticos del modelo de regresión logística con *cutoff* igual a 0.3164

En el cuadro 5.1 se presenta la matriz de confusión y el valor de los estadísticos obtenidos al hacer uso del modelo con el conjunto de validación. Cabe destacar que atendiendo a los valores de la matriz de confusión es posible apreciar que el coste asociado a etiquetar erróneamente un cliente como cliente que abandonará es menor al coste que conlleva etiquetar a un cliente que abandonará la compañía como cliente que va a permanecer. La proporción de los costes

establecidos implica la pérdida de *Precisión* por parte del modelo frente a la ganancia de *Especificidad*. Aproximadamente, 1 de cada 2 clientes etiquetados como *Churn* realmente abandona la compañía y 3 de cada 4 clientes que abandonan son detectados por el modelo.

La figura 5.4 muestra una representación gráfica de la matriz de confusión. Es posible apreciar la distribución de las probabilidades de cada clase (al igual que en la figura 5.1) junto a la frontera determinada por el valor de *cutoff*. El hecho de que los puntos representados en la figura estén esparcidos a lo largo del eje x en cada una de las clases es por un motivo meramente visual.

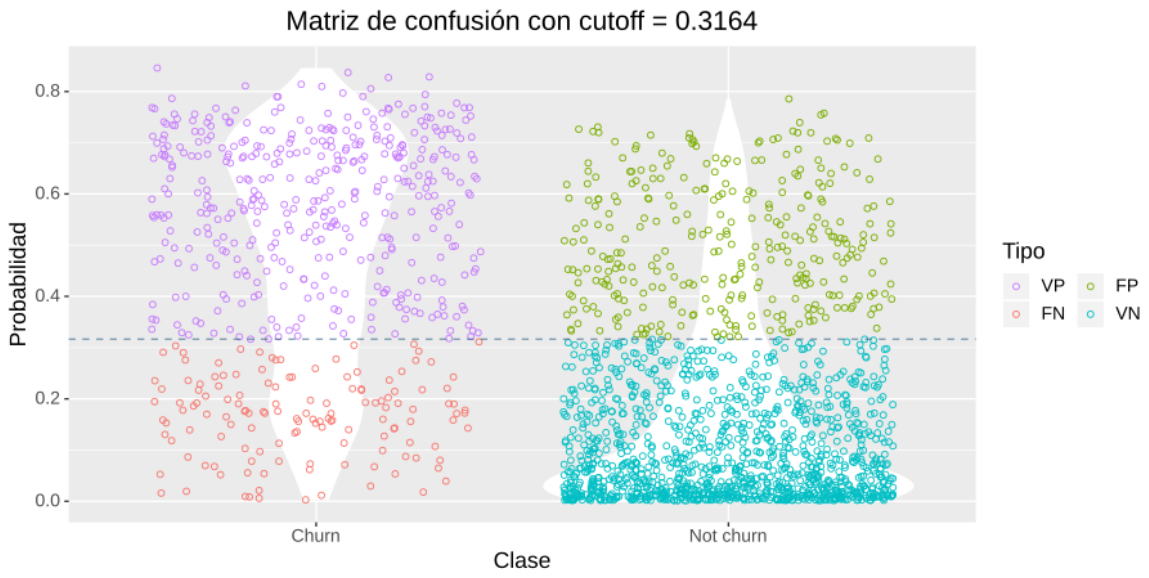

Figura 5.4: Matriz de confusión del modelo de regresión logística

A partir de las tasas de verdaderos positivos y falsos positivos del modelo se obtiene la curva ROC presente en la figura 5.5, el área bajo la curva toma el valor de 0.7681 aproximadamente.

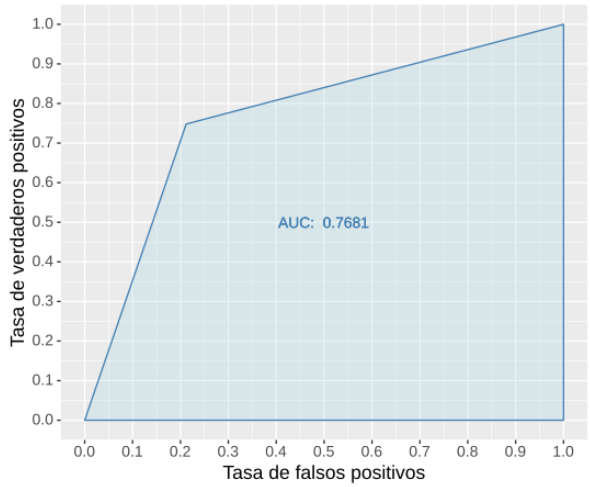

Figura 5.5: Área bajo la curva ROC del modelo de regresión logística y *cutoff* igual a 0.3164

Atendiendo a la información que se presenta en la figura 5.6 es posible observar que existen un conjunto de atributos como *Partner*, *DeviceProtection* u *OnlineBackup* cuya importancia

es muy baja, o lo que es lo mismo, cuyo coeficiente dentro del modelo es cercano a 0. Puede llegar a convenir desechar estos atributos con el objetivo de conseguir un modelo más simple si el comportamiento del mismo no se ve significativamente afectado.

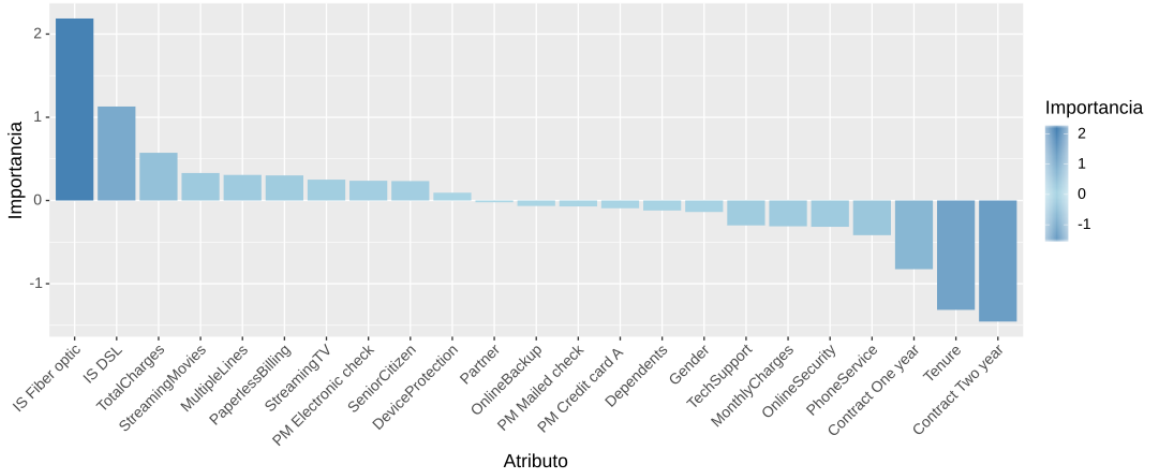

Figura 5.6: Importancia de los atributos del modelo de regresión logística

5.1.1. Selección de características

Tal y como se ha visto anteriormente en esta sección, el modelo de regresión lineal ha sido entrenado a partir del total de los atributos disponibles. Lo cierto es que no todos atributos cobran la suficiente importancia como para que sea necesario tenerlos en cuenta a la hora de predecir el valor del atributo *Churn* en nuevas observaciones. Resulta conveniente intentar simplificar el modelo reduciendo el número de variables de las que depende sin afectar en demasía el comportamiento del mismo. A priori, al usar menos variables se disminuyen los requisitos computacionales y se facilita el uso del modelo pues se ha de disponer de menos datos.

Para la selección de características se optó por utilizar el *Criterio de información de Akaike* (*AIC*) en un proceso iterativo de dos direcciones. *AIC* es un estimador de la calidad relativa de un modelo estadístico dado un conjunto de datos específico [8]. El proceso iterativo se basa en medir la pérdida de información cada vez que se altera el modelo eliminando o añadiendo un atributo [11]. En cada iteración se obtiene el valor del *AIC* de cada modelo resultante al eliminar o añadir (dos direcciones) una de las características disponibles. En la primera iteración no será posible añadir ninguna característica pues el modelo ya incorpora el total de las mismas. Si el *AIC* de alguno de los modelos resultantes es menor que el del modelo actual, este es sustituido (realizándose la adición o eliminación de la característica correspondiente). En el caso de que varios modelos resultantes tengan un *AIC* menor que el del modelo actual se escogerá aquel con mínimo *AIC*. En el momento en el que se han eliminado variables es posible volverlas a incluir en el modelo siempre y cuando el valor del *AIC* disminuya. El valor del *AIC* no tiene ningún significado asociado, únicamente se utiliza para comparar los modelos que se van generando durante el proceso.

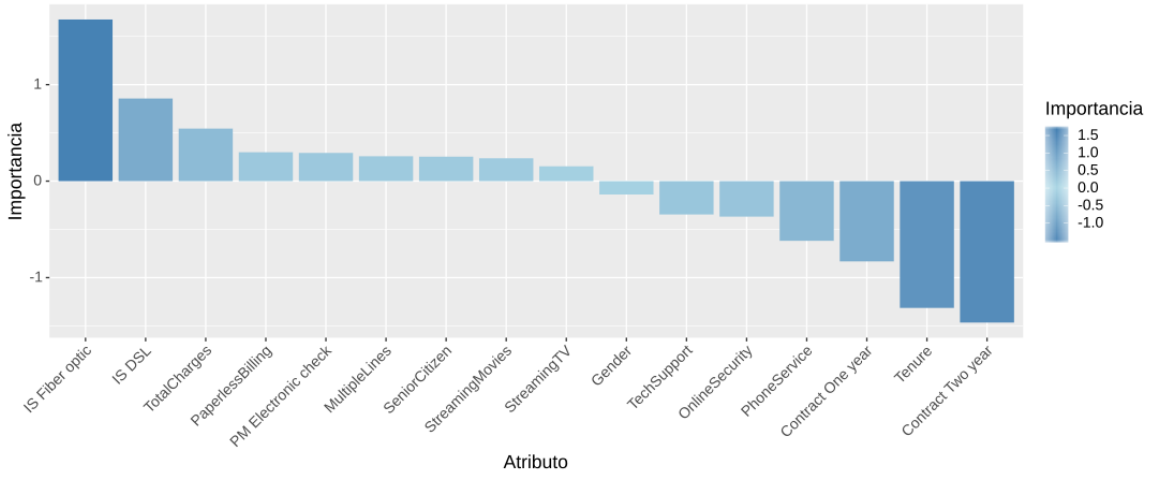

Figura 5.7: Importancia de los atributos del modelo de regresión logística entrenado con selección de características

Tras aplicar el proceso de selección de características se redujo el número de atributos utilizados a 17, los atributos descartados fueron:

- *Partner*
- *MonthlyCharges*
- *PaymentMethod Credit card (A)*
- *DeviceProtection*
- *PaymentMethod Mailed check*
- *Dependents*
- *OnlineBackup*

Resulta interesante apreciar cómo las variables que han sido desechadas son aquellas con menos importancia del modelo de regresión logística entrenado con el total de los atributos. Se corresponden con aquellas situadas hacia el centro de la figura 5.6. También es importante señalar que los coeficientes (importancia) de los atributos (figura 5.7) en el nuevo modelo han sido recalculados y no se corresponden con los mostrados con anterioridad.

Una vez seleccionadas las características, fue necesario comprobar el comportamiento del nuevo modelo simplificado. Para ello se obtuvo el valor de los estadísticos y la matriz de confusión (cuadro 5.2). Cabe señalar que se siguió el mismo procedimiento expuesto anteriormente para dar con el valor de *cutoff* óptimo (0.3016) que minimizase el coste total (759.02 unidades).

	Clase real	
	Churn	*Not churn*
Churn	423	353
Not churn	137	1199

Estadístico	**Valor**
Exactitud	0.7680
Tasa de VP	0.7553
Tasa de VN	0.7725
Precisión	0.5451
Índice Kappa	0.4700
Puntuación F1	0.6332

Cuadro 5.2: Matriz de confusión y estadísticos del modelo de regresión logística entrenado con selección de características y *cutoff* igual a 0.3016

Comparando el valor de los estadísticos y el coste total asociado a la predicción llevada a cabo por el nuevo modelo es posible notar un ligero empeoramiento en su comportamiento. Sin embargo, cabía esperar este empeoramiento pues se ha eliminado un 29 % de los atributos necesarios para predecir si los clientes abandonarán o no. El estudio de la rentabilidad del empeoramiento en el rendimiento frente a la simplificación del modelo debería llevarse a cabo a nivel de negocio cuando el modelo estuviese aplicándose en un entorno real. Por último en las figuras 5.8 y 5.9 pueden observarse la representación gráfica de la matriz de confusión y la curva ROC respectivamente.

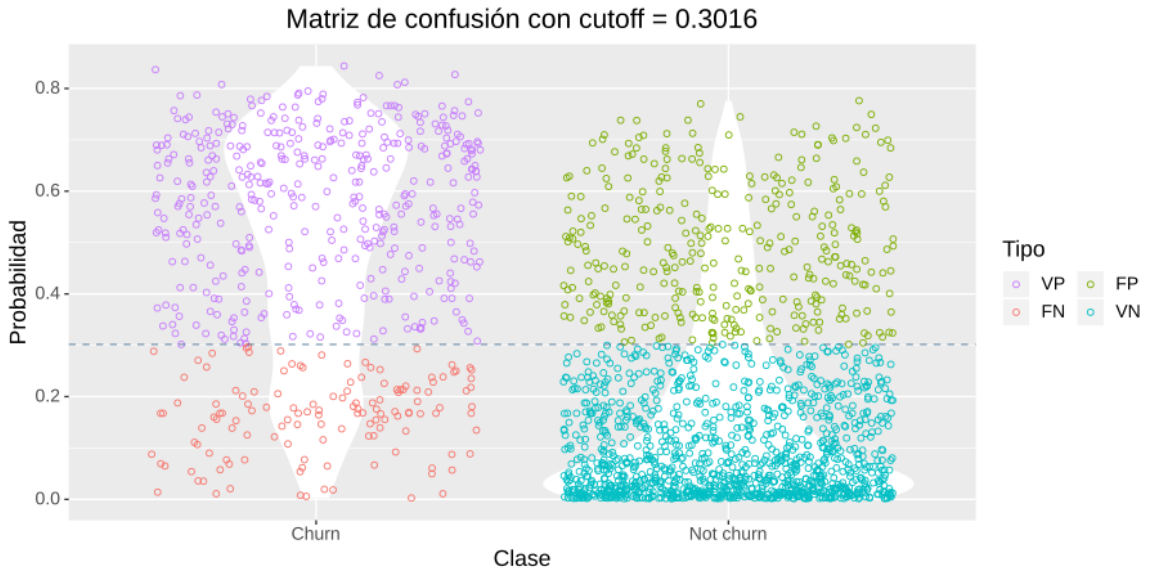

Figura 5.8: Matriz de confusión del modelo de regresión logística entrenado con selección de características

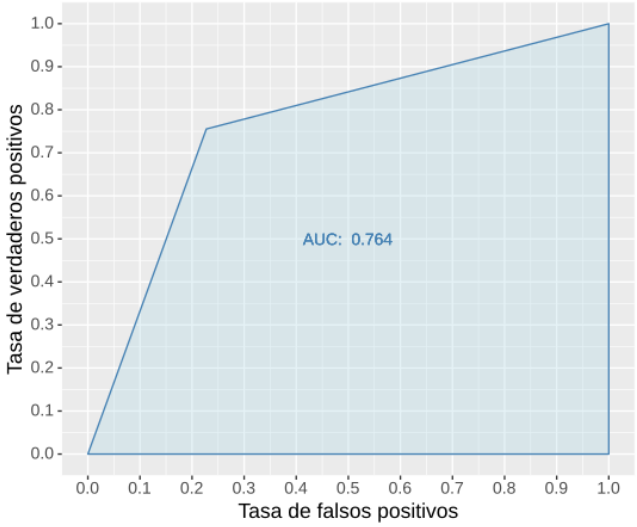

Figura 5.9: Área bajo la curva ROC del modelo de regresión logística entrenado con selección de características y *cutoff* igual a 0.3016

5.2. Árbol de decisión

Un árbol de decisión es un modelo de clasificación creado a partir de la división de los datos en subconjuntos. El nombre de este tipo de modelos se debe a que las divisiones del conjunto de datos van creando una estructura en forma de árbol en la que cada nodo interno simboliza una pregunta sobre un atributo, cada rama una respuesta a la pregunta del nodo interno previo y los nodos hoja una clase del atributo categórico que se intenta predecir. Durante la construcción del modelo, la división de los datos se hace teniendo en cuenta el concepto de homogeneidad. La decisión de qué atributo seleccionar para la próxima división se toma atendiendo a la homogeneidad de los subconjuntos resultantes. La homogeneidad de los subconjuntos puede medirse atendiendo a la entropía, el índice de *Gini* o la ganancia de información.

Para el entrenamiento del modelo se optó por subdividir el número de observaciones del conjunto de datos en dos grupos teniendo en cuenta la proporción de los distintos valores del atributo de clase: entrenamiento y validación. El conjunto de entrenamiento se conformó a partir del 70 % de las observaciones disponibles y el de validación con el 30 % restante. Durante la generación del árbol de decisión se hizo uso de validación cruzada de 5 pliegues sobre el conjunto de entrenamiento. Los atributos seleccionados para la construcción del modelo fueron:

- Atributos categóricos originales y categóricos simplificados (sección 3.3)
- Atributos numéricos

La figura 5.10 presenta el árbol de decisión obtenido como primera aproximación. Atendiendo al árbol obtenido, se puede concluir que únicamente existe riesgo de abandono cuando el cliente bajo estudio tiene un contrato de tipo mes a mes, servicio de conexión a Internet por fibra óptica y lleva menos de 15 meses en la compañía. A priori, no parece que este modelo de clasificación vaya a tener un buen comportamiento.

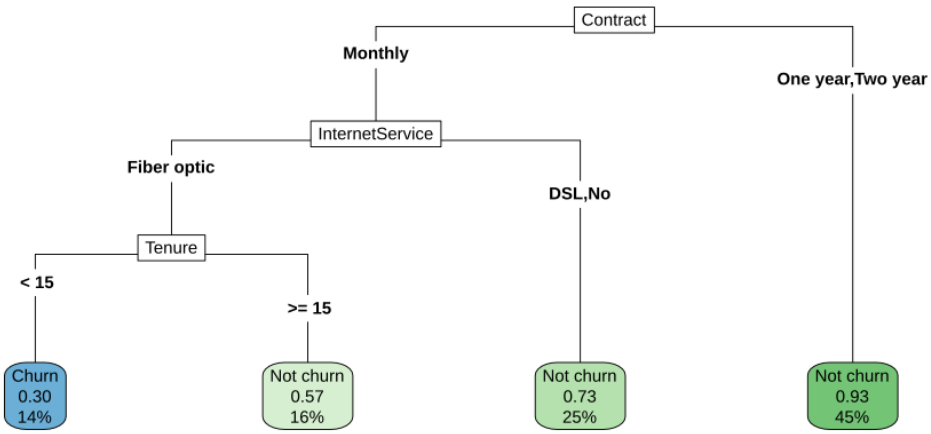

Figura 5.10: Árbol de decisión

De los valores que se muestran en la matriz de confusión y la tabla de los estadísticos del cuadro 5.3 se puede concluir que el rendimiento del clasificador está muy afectado por el problema de clases desbalanceadas del conjunto de datos. El clasificador no consigue detectar las observaciones etiquetadas como *Churn* como debiera. Que la *Tasa de VP* sea tan baja y la *Precisión* relativamente alta indica que el clasificador únicamente está prediciendo correctamente los casos más "sencillos" o notables de clientes que van a abandonar la compañía. Aproximadamente, de cada 10 clientes que abandonan la compañía, únicamente 3 son detectados por el modelo. El área bajo la curva ROC del clasificador (visible en la figura 5.11) es próxima a la de un clasificador completamente aleatorio.

	Clase real	
	Churn	*Not churn*
Churn	204	91
Not churn	356	1461

Estadístico	**Valor**
Exactitud	0.7883
Tasa de VP	0.3643
Tasa de VN	0.9414
Precisión	0.6915
Índice Kappa	0.3601
Puntuación F1	0.4772

Cuadro 5.3: Matriz de confusión y estadísticos del árbol de decisión

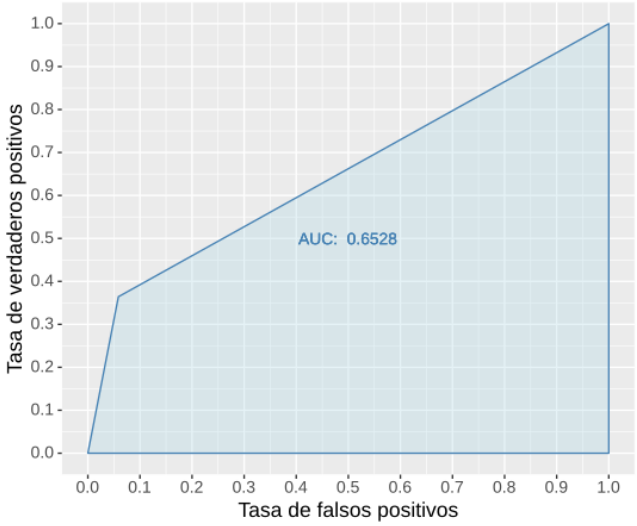

Figura 5.11: Área bajo la curva ROC del árbol de decisión

A modo de solución, se creyó conveniente especificar una matriz de costes, al igual que en el caso de la búsqueda del mejor *cutoff* en el modelo de regresión logística. De esta manera es posible guiar al clasificador dándole prioridad a la correcta clasificación de la clase menos representada, que además es la que más interesa desde el punto de vista del problema. Se decidió asignar los mismos costes que en el modelo de regresión logística para los dos tipos de fallos de clasificación, es decir, 3 unidades para los falsos negativos y 1 unidad para los falsos positivos. Tomando estos valores, el coste total asociado a la predicción realizada por el árbol de decisión previo es de 1159 unidades, muy por encima de los obtenidos por los clasificadores de regresión logística estudiados.

Especificando la matriz de costes, el árbol obtenido es el que se presenta en la figura 5.12. Resulta interesante comentar las reglas de clasificación que se extraen de la lectura del árbol,

por ejemplo, aquellos clientes con contratos de 1 o 2 años serán clasificados como *Not churn* independientemente del resto de atributos. Al construirse el modelo, el 11.4 % de clientes que abandonaron la compañía durante el último mes teniendo este tipo de contratos (figura 4.6) no fue suficiente como para considerar crear una ramificación extra en la que se considerase otro atributo para determinar la categoría de la observación. Del árbol también se puede extraer la conclusión de que aquellos clientes con contrato mensual y servicio de conexión a Internet por fibra óptica abandonarán la compañía. Como puede intuirse, independientemente del valor de los estadísticos de la predicción sobre el conjunto de validación, las reglas de clasificación que extraen del árbol no parecen ser muy congruentes.

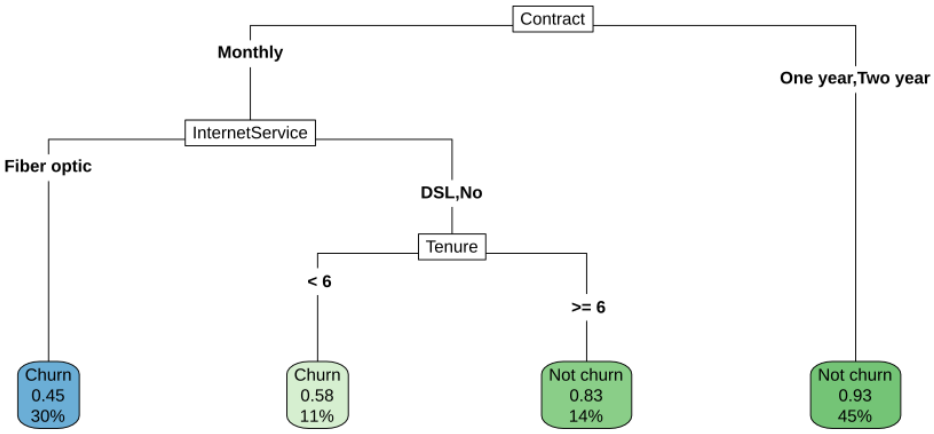

Figura 5.12: Árbol de decisión obtenido especificando una matriz de costes

Para valorar el resultado de incluir la matriz de costes en el proceso de construcción del clasificador es necesario estudiar el nuevo valor que toman los estadísticos. El cuadro 5.4 muestra estos valores y la matriz de confusión asociada a la predicción sobre el conjunto de validación. Como cabría esperar, el comportamiento del clasificador ha mejorado sustancialmente. Especificando un coste mayor a los falsos negativos se ha conseguido aumentar la *tasa de verdaderos positivos* a costa de empeorar la *Precisión*. El coste asociado a la predicción realizada por el nuevo árbol de decisión es de 828 unidades, bastante inferior al obtenido por el árbol de decisión sin matriz de costes.

	Clase real	
	Churn	*Not churn*
Churn	428	432
Not churn	132	1120

Estadístico	**Valor**
Exactitud	0.7329
Tasa de VP	0.7643
Tasa de VN	0.7216
Precisión	0.4977
Índice Kappa	0.4149
Puntuación F1	0.6028

Cuadro 5.4: Matriz de confusión y estadísticos del árbol de decisión obtenido especificando una matriz de costes

La figura 5.13 muestra que el área bajo la curva ROC también ha aumentado significativamente. En definitiva y como cabría esperar, al solucionar el problema de desbalanceo de clases en el conjunto de datos se ha logrado una mejora en el rendimiento del modelo, consiguiendo un comportamiento similar al del modelo de regresión logística. Sin embargo, de la visualización del árbol de decisión se han podido interpretar reglas de clasificación que en principio, no son muy coherentes.

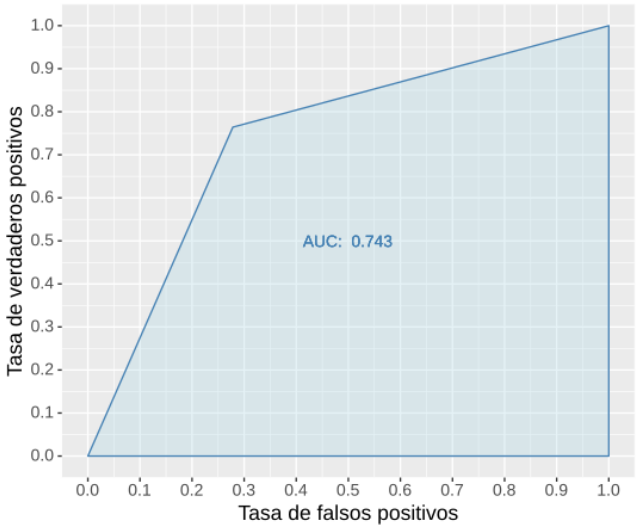

Figura 5.13: Área bajo la curva ROC del árbol de decisión obtenido especificando una matriz de costes

5.3. *Naïve* Bayes

Los clasificadores *naïve* Bayes son una familia de clasificadores probabilísticos simples basados en el uso del teorema de Bayes asumiendo "ingenuamente" la independencia total de los sucesos (características o atributos) bajo estudio, es decir, se presupone que el valor que toma un atributo es independiente al valor de cualquier otro dada la variable de clase, ignorando cualquier posible correlación existente. Otra característica importante de este tipo de clasificadores consiste en que se asume que cada uno de los atributos a los que se atiende durante la clasificación tienen el mismo efecto o peso para la predicción de la clase dada una observación. Para cada valor posible de cada atributo se estima la probabilidad condicionada a cada una de las clases de la variable de clase atendiendo a las frecuencias de aparición en los ejemplos del conjunto de entrenamiento del modelo. Esto introduce un problema ya que el buen comportamiento del clasificador depende completamente de la representatividad de los datos presentes en el conjunto de entrenamiento.

Dada una observación, se estimará la probabilidad de que la misma pertenezca a cada una de las posibles categorías de la variable de clase atendiendo a los valores de los demás atributos. La observación se etiquetará considerando la clase con mayor probabilidad estimada.

Al igual que con los clasificadores generados anteriormente, se escogió dividir el conjunto de datos en dos subconjuntos: entrenamiento y validación. El conjunto de entrenamiento supone un 70 % del total y el de validación el 30 % restante. En cuanto a las variables usadas para la generación del clasificador, se optó por hacer uso únicamente de variables categóricas:

- Atributos categóricos originales
- Atributos numéricos categorizados

Realizando la predicción sobre el conjunto de validación se obtiene la matriz de confusión que se presenta en la figura 5.5. Con el objetivo de poder estudiar el coste total asociado a la predicción realizada, se asignaron costes a los falsos negativos (3 unidades) y falsos positivos (1 unidad), al igual que en el caso de los modelos anteriores. Atendiendo al valor de los estadísticos, la *tasa de falsos negativos* toma un valor demasiado alto. Esto es totalmente indeseable pues simboliza aquellos clientes que abandonan la compañía sin ser detectados por el clasificador. Por el contrario, la *tasa de falsos negativos* toma un valor relativamente bajo y aceptable. Conociendo las características de los clasificadores *naïve* Bayes (basados en la representatividad del conjunto de entrenamiento), es posible deducir que el problema de clases desbalanceadas que presenta el conjunto de datos esté pasando factura al rendimiento del clasificador.

	Clase real	
	Churn	*Not churn*
Churn	386	303
Not churn	174	1249

Estadístico	**Valor**
Exactitud	0.7741
Tasa de VP	0.6893
Tasa de VN	0.8048
Precisión	0.5602
Índice Kappa	0.4602
Puntuación F1	0.6181

Cuadro 5.5: Matriz de confusión y estadísticos del clasificador *naïve* Bayes

El coste total asociado a la predicción realizada sobre los datos de validación es de 825 unidades y el área bajo la curva ROC (visible en la figura 5.14) es 0.747 aproximadamente.

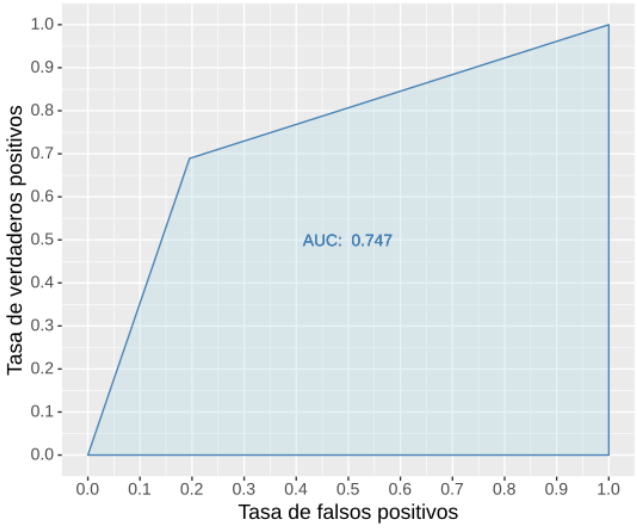

Figura 5.14: Área bajo la curva ROC del clasificador *naïve* Bayes

Capítulo 6

Comparativa de los modelos

En este capítulo se presenta una comparativa del comportamiento o bondad de los clasificadores presentados en el capítulo 5. Para cada clasificador, se compararán el valor de los estadísticos, el área bajo la curva ROC y el coste total asociado a la predicción realizada sobre el conjunto de validación. Para el coste total se han especificado los valores de 3 unidades para los falsos negativos y 1 unidad para los falsos positivos. En concreto, los modelos a comparar son los siguientes:

- Clasificador de regresión logística con selección de características (RL)
- Árbol de decisión con matriz de costes (AD)
- Clasificador *naïve* Bayes (NB)

En el cuadro 6.1 se exponen los valores de las métricas obtenidas para cada clasificador. Resulta interesante apreciar cómo todos los valores de cada métrica son relativamente cercanos entre los distintos clasificadores.

	Exactitud	*VP*	*VN*	*Precisión*	*Kappa*	*F1*	*AUC*	*Coste total*
RL	0.7680	0.7553	0.7725	0.5451	0.47	0.6332	0.764	759.02
AD	0.7329	0.7643	0.7216	0.4977	0.4149	0.6028	0.743	828
NB	0.7741	0.6893	0.8048	0.5602	0.4602	0.6181	0.747	825

Cuadro 6.1: Tabla comparativa de clasificadores

A continuación se presentan cuadros con clasificaciones de los clasificadores atendiendo al valor que toma cada una de las métricas bajo estudio.

	Exactitud	Posición
RL	0.7680	#2
AD	0.7329	#3
NB	0.7741	#1

	VP	Posición
RL	0.7553	#2
AD	0.7643	#1
NB	0.6893	#3

	VN	Posición
RL	0.7725	#2
AD	0.7216	#3
NB	0.8048	#1

Cuadro 6.2: Clasificación de clasificadores atendiendo a los estadísticos *Exactitud*, *Tasa de VP* y *Tasa de VN*

	Precisión	Posición
RL	0.5451	#2
AD	0.4977	#3
NB	0.5602	#1

	Kappa	Posición
RL	0.47	#1
AD	0.4149	#3
NB	0.4602	#2

	F1	Posición
RL	0.6332	#1
AD	0.6028	#3
NB	0.6181	#2

Cuadro 6.3: Clasificación de clasificadores atendiendo a los estadísticos *Precisión*, *Índice Kappa* y *Puntuación F1*

	AUC	Posición
RL	0.764	#1
AD	0.743	#3
NB	0.747	#2

	Coste total	Posición
RL	759.02	#1
AD	828	#3
NB	825	#2

Cuadro 6.4: Clasificación de clasificadores atendiendo al área bajo la curva ROC y al coste total asociado a la predicción realizada sobre el conjunto de validación

En el cuadro 6.5 es posible encontrar una clasificación global en la que se le asigna una puntuación a cada clasificador atendiendo a las posiciones de cada una de las clasificaciones individuales presentadas anteriormente. Por cada primera posición se asignan 3 puntos, por cada segunda posición se asignan 2 puntos y por cada tercera posición se asigna 1 punto. Atendiendo a la información del cuadro, el clasificador con mayor puntuación es el de regresión logística, quedando en primer lugar en todas las clasificaciones individuales de las métricas que agrupaban más de un estadístico (*Índice Kappa*, *Puntuación F1*, *AUC* y *Coste total*).

	1ª posición	2ª posición	3ª posición	Puntuación	Posición global
RL	4	4	0	20	#1
AD	1	0	7	10	#3
NB	3	4	1	18	#2

Cuadro 6.5: Clasificación global de los clasificadores

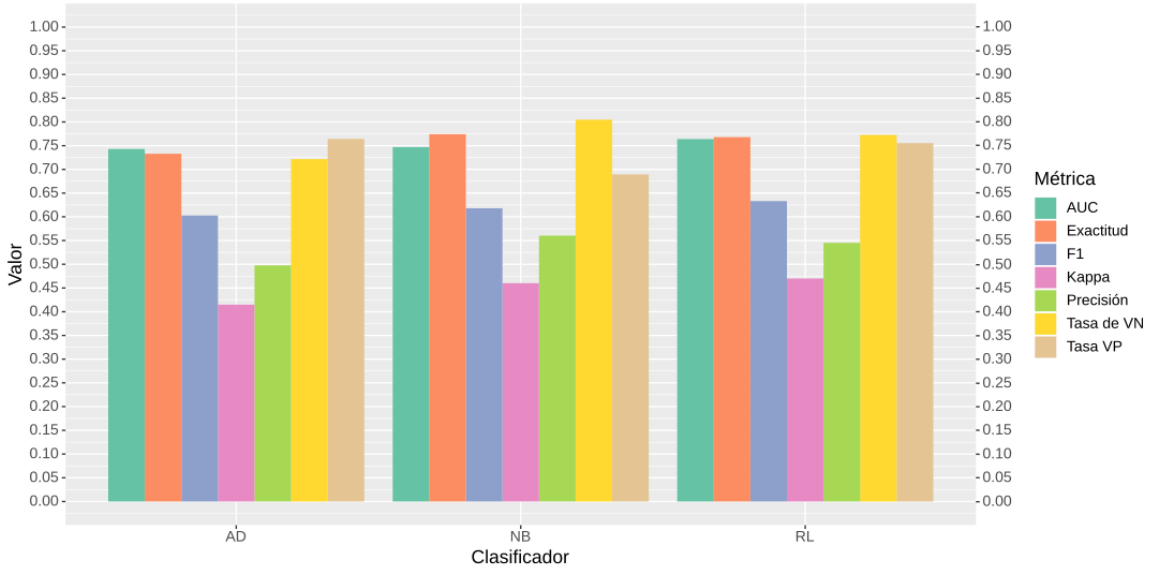

Figura 6.1: Valores de las métricas para cada clasificador

Por último, la figura 6.1 muestra una comparación gráfica de cada una de las métricas de cada clasificador cuyos valores quedan comprendido en el rango [0, 1] y en la figura 6.2 puede encontrarse la comparación de la métrica *Coste total*.

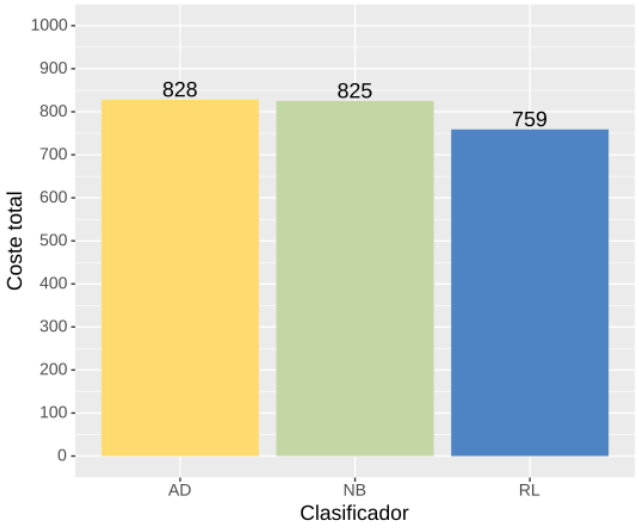

Figura 6.2: Coste total de las predicciones de cada clasificador

Capítulo 7

Conclusiones

Atendiendo a los clasificadores que se han comparado en el capítulo 6 de la memoria, el que **mejor comportamiento** ha demostrado sobre el conjunto de validación es el **clasificador de regresión logística**. Además de ocupar el primer puesto de la clasificación global de clasificadores(visible en el cuadro 6.5), parece ser el más adecuado para el problema enfrentado. El hecho de que se asocie un coeficiente o peso a cada una de las variables bajo estudio y de que sobre el total de aportaciones se delibere si una observación pertenece a una u otra clase parece más fiable, robusto y preciso que tomar la decisión atendiendo a reglas de clasificación con poca coherencia o sentido aparente (extraídas del árbol de decisión, expuestas en la sección 5.2) o asumiendo la independencia total de los atributos y que los datos de entrenamiento son altamente representativos (como en el caso del clasificador *naïve* Bayes, sección 5.3).

A continuación se presentan una serie de posibles líneas futuras para la continuación de la resolución del problema de prevención de cancelación de contratos en compañías de telecomunicaciones:

- Si bien es cierto que la pérdida de cualquier tipo de clientes es extremadamente dañina para las compañías de telecomunicaciones, el intento de priorizar la retención de clientes con altas cuantías mensuales puede resultar interesante de cara al beneficio económico.
- La búsqueda de un mejor clasificador siempre supondrá una alternativa. En concreto, resultaría conveniente estudiar el comportamiento de los *ensemble methods*, basados en la combinación de múltiples modelos denominados "débiles" para obtener un modelo más robusto y preciso.
- La selección de características es una posibilidad poco explorada a lo largo del proyecto en la que se ha de ahondar con el objetivo de conseguir mejores modelos de clasificación.
- El estudio de otras variables de diversa índole también debería tenerse en cuenta, cabe destacar que los datos se han extraído de una plataforma abierta. Con total probabilidad, la compañía de telecomunicaciones cuenta con muchos más atributos a los que prestar atención durante la resolución del problema.

En conclusión, la calidad de los clasificadores generados demuestra que es posible hacer frente al problema de prevención de cancelación de contratos en compañías de telecomunicaciones. A pesar de que los clasificadores que se han construido para este proyecto no son infalibles ni tienen un comportamiento extremadamente bueno, sí que se ha podido evidenciar que son capaces de identificar un alto porcentaje de los clientes que van a abandonar la compañía. Cierto es que se debería verificar la calidad de los clasificadores realizando pruebas en un entorno real

controlado con el propósito de justificar si es posible hacer uso de alguno de estos clasificadores como mecanismo de prevención y retención de clientes. En caso afirmativo y asumiendo que al hacer un uso controlado del mismo solo pueden obtenerse beneficios, el clasificador que se utilizase supondría una gran ventaja en cuanto a la consecución del objetivo de minimizar el abandono por parte de los clientes y en consecuencia, en cuanto a la competitividad de la compañía que lo implantase. Como cabría esperar, dicha herramienta debería ser estudiada, analizada y actualizada continuamente con el propósito de asegurar su buen comportamiento y adaptabilidad, y más teniendo en cuenta la naturaleza cambiante del problema y de los datos. Uno de los factores más importantes a tener en cuenta a la hora de realizar un estudio de la viabilidad de la implantación del clasificador en un entorno real es el coste asociado a cada uno de los posibles errores de clasificación. En este proyecto se ha asignado el triple de coste a un falso negativo (cliente clasificado como *Not churn* que termina cancelando su contrato) que a un falso positivo (cliente clasificado como *Churn* que no tiene intención real de abandonar la compañía). Esta decisión fue tomada únicamente con el propósito de dar un valor de ejemplo, aunque bien es cierto que los más probable es que la proporción de coste entre los tipos de error sea incluso mayor.

Bibliografía

[1] BlastChar. Kaggle telco customer churn dataset. https://www.kaggle.com/blastchar/telco-customer-churn. [Online 26-Diciembre-2019].

[2] Codeacademy. Normalization. https://www.codecademy.com/articles/normalization. [Online 30-Diciembre-2019].

[3] Norman R Draper and Harry Smith. *Applied regression analysis*, volume 326. John Wiley & Sons, 1998.

[4] F Reichheld Frederick and W Earl Sasser. Zero defections: quality comes to services. *Harvard Business Review*, 68(5):105, 1990.

[5] Rohan Gupta. An introduction to discretization techniques for data scientists. https://towardsdatascience.com/an-introduction-to-discretization-in-data-science-55ef8c9775a2. [Online 28-Diciembre-2019].

[6] Kaggle Inc. Kaggle: Your home for data science. https://www.kaggle.com/. [Online 26-Diciembre-2019].

[7] Ming-Yu Liu. Choosing logisitic regression's cutoff value for unbalanced dataset. http://ethen8181.github.io/machine-learning/unbalanced/unbalanced.html. [Online 04-Enero-2020].

[8] Richard McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC, 2018.

[9] Faraz Rahman. Telco customer churn - logistic regression. https://www.kaggle.com/farazrahman/telco-customer-churn-logisticregression. [Online 26-Diciembre-2019].

[10] Pavan Raj. Telecom customer churn prediction. https://www.kaggle.com/pavanraj159/telecom-customer-churn-prediction. [Online 26-Diciembre-2019].

[11] Ashutosh Tripathi. What is stepaic in r? https://medium.com/@ashutosh.optimistic/what-is-stepaic-in-r-a65b71c9eeba. [Online 04-Enero-2020].

[12] Wikipedia. Logit. https://en.wikipedia.org/wiki/Logit. [Online 03-Enero-2020].