Taller 002 grupal

Modelo de clasificación para predecir la deserción en la industria móvil colombiana.

Daniel Fernando Neira Pardo[a,c] , Jhoan Esteban Romero Muñoz[a,c] , Laura Juliana Mora Páez[a,c], Robinsson Yesid Sánchez Deantonio[a,c],

Ana María Beltrán Cortés[b,c]

[A]Maestría en Analítica para la Inteligencia de Negocios
Maestría en Banca y Finanzas
[b]Profesor, Departamento de Ingeniería Industrial
[c]Pontificia Universidad Javeriana, Bogotá, Colombia

1. ENTENDIMIENTO DEL NEGOCIO

Mientras que en América Latina la penetración de la telefonía móvil está en niveles cercanos al 70%, en Colombia, el último trimestre de 2021 cerró con 75 millones de abonados (usuarios de una red pública de telecomunicación celular), equivalente a una penetración de mercado del 143%, valor que duplica el benchmark de la región y nos ubica en tercer lugar detrás de Brasil y Argentina. Dicho comportamiento es una clara muestra de la importancia de este mercado en el territorio nacional y de lo avanzados que estamos frente a otros mercados semejantes.

Desde el año 2011 se ha dado un crecimiento sostenido en términos de abonados como lo muestra la siguiente gráfica:

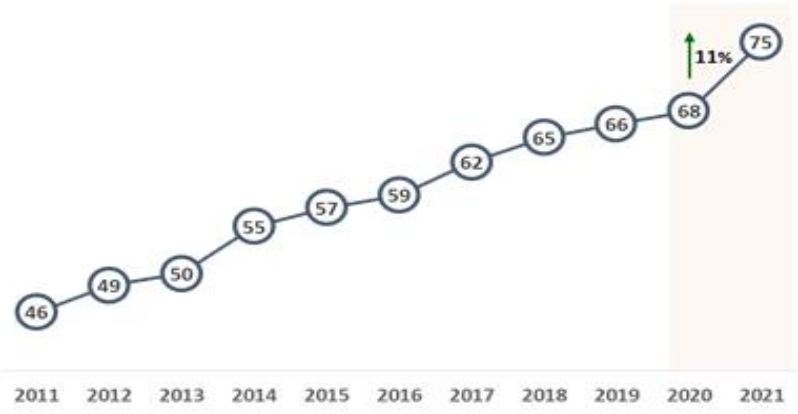

Figura 1. Crecimiento de abonados en el mercado

Fuente: https://postdata.gov.co/dataset/abonados-ingresos-y-tr%C3%A1fico-de-telefon%C3%ADa-m%C3%B3vil

Mientras que los crecimientos de abonados en 2019 y 2020 se encontraban en niveles del 2%, 2021 presentó el incremento más importante en la historia, desde que se hace seguimiento a este indicador al situarse en el 11%. Vale la pena resaltar que dicho crecimiento se dio gracias a la coyuntura creada por la pandemia, donde el uso de este medio de comunicación se hizo más importante.

Aunque es notorio el crecimiento de usuarios, la entrada de nuevos jugadores y la feroz competencia entre los principales competidores ha impactado en el ingreso del sector. En el segundo semestre de 2021, los ingresos por voz presentaron una disminución del 7.3% y los ingresos por mensaje de texto cayeron un 3.3% en comparación con el mismo periodo del año anterior.

Tabla 1: Participación Mercado Telefonía Móvil IV Trimestre 2021

Operador	Abonados	% Part.
Claro	35.061.951	47%
Movistar	18.777.096	25%
Tigo	14.517.046	19%
Virgin	2.734.135	4%
WOM	1.951.232	3%
Otros	2.014.651	3%
Total general	75.056.111	100%

En Colombia, existen 5 grandes jugadores en la industria celular que se reparten el 97% de la participación del mercado, siendo Claro el más importante con casi la mitad de la cuota total. Uno de los operadores que vale la pena hacer mención es WOM, ya que entro a competir en el mercado en el año 2021 y logró ocupar el quinto lugar en el escalafón, por encima de competidores de mayor antigüedad en el mercado como lo es el caso de Avantel.

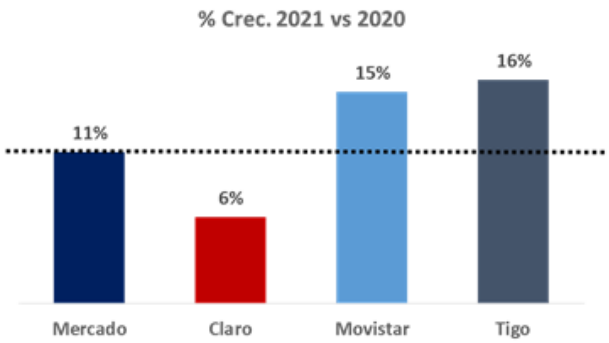

Figura 2. Crecimiento de competidores en el mercado

Si bien la dinámica de crecimiento del mercado ha sido muy favorable, no todos los competidores han presentado un incremento a la par del mercado, Movistar y Tigo han presentado crecimiento por encima del promedio, mientras que Claro presenta un crecimiento del 6%, considerablemente por debajo del global y sus competidores más

cercanos. A pesar de que los abonados de Claro no han disminuido, se ha cedido participación en el mercado.

La razón principal por la que Claro presenta este comportamiento inusual se debe a la deserción, una cantidad significativa de sus abonados se ha trasladado y adquirido servicios con sus competidores. Es por esto que para la empresa se hace imperativo controlar el indicador de deserción o *churn* teniendo en cuenta que, según un análisis realizado por la empresa, el costo de mantener o fidelizar a un cliente es una quinta parte del costo asociado a que este cancele los servicios contratados. El éxito de la implementación del modelo está en minimizar el indicador de churn, mediante la implementación de una herramienta de predicción que permita generar alertas tempranas que permita implementar estrategias de retención.

Objetivos de negocio:

- Identificar posibles clientes en riesgo de deserción del servicio.
- Diseñar estrategias que permitan mejorar la interacción con los clientes, de forma que se minimice el churn.
- Obtener un kpi de ganancia neta positivo.

Criterios de éxito del negocio:

- Obtener variables ayuden a comprender el comportamiento de los clientes previo a la deserción.
- Obtener los datos para la comparación entre atraer nuevos clientes y retener los existentes.

Objetivos de minería:

- Identificar variables que sean influyentes a la hora de determinar si un cliente desertara o no.

Criterios del proceso de minería de datos:

- Obtener un AUC superior al planteado por el ingenio de la plataforma Kaggle, es decir superior a 0.68208.

2. ENTENDIMIENTO DE LOS DATOS

Se cuenta con una base de datos de 8.243 registros de los usuarios de la empresa Claro, y una base con los datos de 2.000 registros, en las cuales se debe realizar un pronóstico de deserción del cliente. Si bien se desconoce el proceso de recolección de los datos, se conoce que las bases de datos fueron entregadas en dos archivos de Excel “traintelco.xlsx”, donde se encuentran los 8.243 registros y los 11 atributos iniciales, y el otro archivo “testelco.xlsx” con 2.000 registros, donde se debe determinar si el cliente abandone la compañía. Cabe resaltar que los sets de datos cuentan con una estructura de tabla clásica, adicionalmente, no se cuenta con datos referentes a información geográfica, sin embargo, se sabe que los datos corresponden al consumo

durante los últimos seis meses del cliente con el operador para el periodo de tiempo de 2018-2019.

A continuación, se presenta el diccionario de variables en la Tabla 2, con la descripción de cada uno de los atributos con los que se cuenta en ambas bases de datos, con el fin de dar mayor claridad sobre estos.

Tabla 2: Diccionario de variables.

Diccionario de variables			
Variable	**Explicación**	**Tipo de dato**	**Valores posibles**
id	Identificador anónimo y único del cliente registrado en el sistema.	Numérico / Cualitativa	5- 18310
Fecha de nacimiento	Fecha de nacimiento del cliente, dd/mm/yyyy.	Date/ Cuantitativa	1940/09/10 08:22:08 – 1992/12/05 04:14:37
tipo cliente	Identificador de la categoría de cliente.	Numérica/ Cualitativa	1 - Hombre, 2 - Mujer, 3 - Cliente empresarial.
Factura online	Indica si el cliente recibe únicamente su factura online.	Numérico /Binaria/ Cualitativa	0-1
Antigüedad equipo	Meses de antigüedad del equipo.	Numérico/ Cuantitativa	1-48
Plan de datos	Indica si el cliente tiene plan premium de datos.	Numérico /Binaria/ Cualitativa	0-1
Facturación	Total facturación de los últimos seis meses (suma).	Numérico **/** Cuantitativa	-112588 - 626111
Mora	Días de mora acumulados.	Numérico **/** Cuantitativa	0 - 118
Fecha inicio contrato	Fecha de inicio de su primer contrato de plan pospago en la compañía.	Date/ Cuantitativa	2011/12/13 07:49:29 – 2019/01/03 07:52:55
minutos	Total minutos consumidos en los últimos seis meses.	Numérico/ Cuantitativa	100 - 14103
resultado	Registro del churn (abandono del cliente)	Numérico /Binaria/ Cualitativa	0-1

Con el fin de explorar los datos, tanto de train como de test, y tener una primera impresión de cuales datos podían influenciar más a la hora de generar las predicciones, se inició por una descripción de los atributos de carácter cuantitativo, en donde se puede observar en la Tabla 3 que los atributos de facturación, antiguedad_equipo y mora, presentan kurtosis cercanas a 0, mientras que el atributo minutos presenta una kurtosis de 54.80, demostrándonos una posible asimetría en los datos, por su parte las desviaciones estándar de antiguedad_equipo y mora son cercanas a 0, debido a que

sus datos se encuentran agrupados alrededor de la media, sin embargo, para minutos y más específicamente facturación se observan desviaciones estándar altas, indicando que los valores que pueden tomar los datos se extienden por un rango bastante amplio. Adicionalmente, para el caso de facturación se puede empezar a observar una presencia de datos atípicos ya que el valor mínimo que puede tomar este atributo es de -112.588.

Tabla 3: Medidas estadísticas de las variables numéricas

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
antiguedad_equipo	1	8243	24.27	15.15	24	24.22	22.24	1	48	47	0.02	-1.63	0.17
facturacion	2	8243	267765.69	95596.02	268063	267903.86	96134.75	-112588	626111	738699	-0.02	0.01	1052.92
mora	3	8243	18.13	21.03	6	15.27	8.90	0	118	118	0.91	0.10	0.23
minutos	4	8243	1183.82	932.84	1103	1102.78	585.63	100	14103	14003	6.01	54.80	10.27

Posteriormente, se realizan histogramas con el fin de observar el comportamiento de los atributos numérico-cualitativos (Fig 3), para el caso de tipo cliente se observa que la mayoría de los clientes pertenecen al tipo 1 (hombre) y 2 (mujer); adicionalmente, si se compara la cantidad de clientes tipo 3 con los dos anteriores se destaca por tener pocos; por su parte existen menos clientes que obtienen su factura online que los que no, sin embargo, no se presenta una diferencia muy amplia. Finalmente, en el caso del atributo de plan de datos se observa que la gran mayoría de los usuarios cuentan con este.

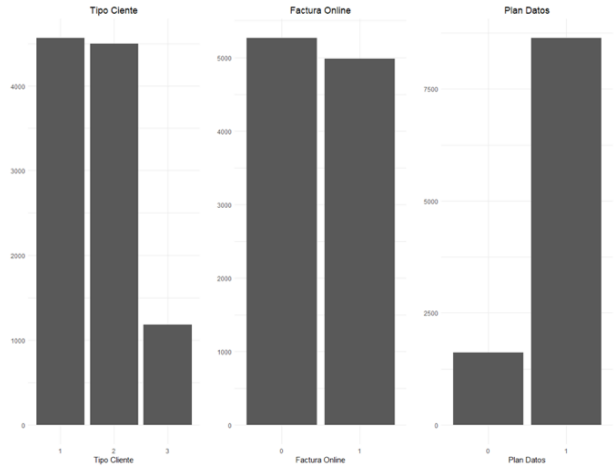

Figura 3. Distribución de variables categóricas

Por otro lado se realizaron boxplot para observar de manera gráfica el comportamiento de los atributos y sus interacciones con la variable a predecir (deserción), según la clasificación de las variables categóricas que se utilizan, como se muestra en la Figura 4 donde se destaca una alta presencia de atípicos en las variables de Facturación, Mora y Minutos, sin importar bajo cual variable estaba categorizada, sin embargo, para el caso de Antigüedad del equipo al ser agrupada por tipo_cliente y comparada contra la deserción no se muestra presencia de datos atípicos, mientras que cuando es agrupada por factura_online si se observa una alta presencia de datos atípicos en las diferentes combinaciones; por su parte al ser agrupada por plan_datos solo se observa una presencia de datos atípicos en las combinaciones donde los clientes no cuentan con un plan de datos.

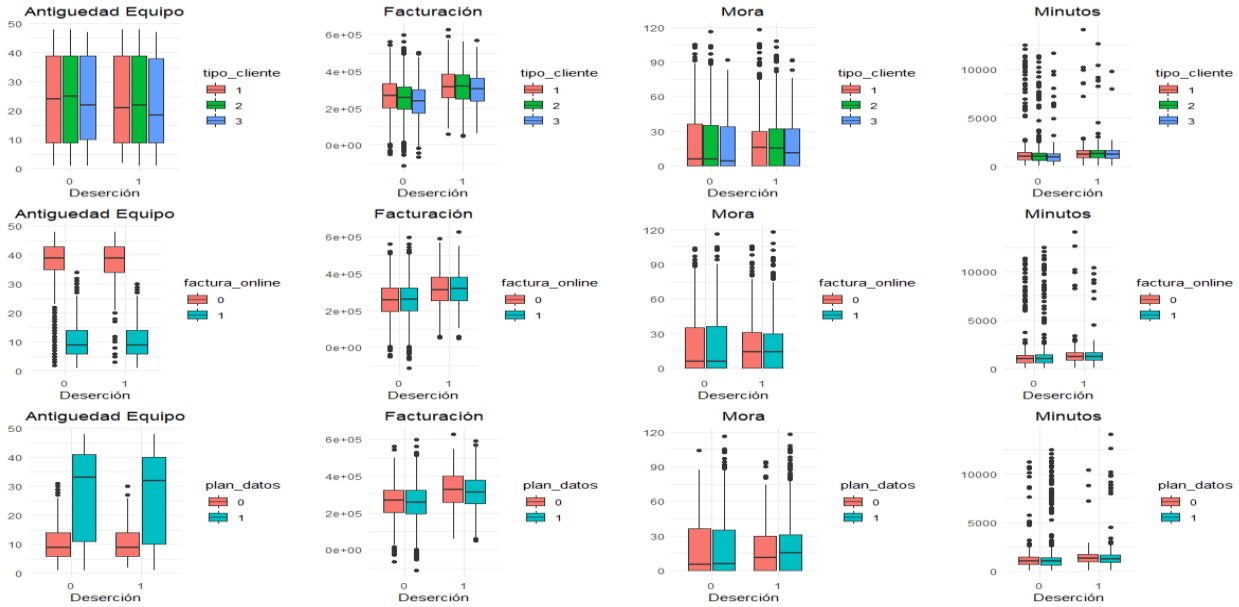

Figura 4. Boxplot de variables numéricas diferenciado por variables categóricas

3. PREPARACIÓN DE LOS DATOS

Los ejercicios se realizan tanto para la base de train, cómo para la de test.

Notamos que la base tiene un desbalance en cantidades de 0 y 1, se evidencia necesidad de balanceo, para la variable dependiente (resultado), según el summary obtenido en R-studio, tenemos que para respuesta 0 fueron 6.882 (83,49%) registros, y para respuesta 1 fueron 1.361 (16,51%). Para solucionar este problema, se realizan pruebas con la metodología de oversampling para tener en los porcentajes de respuesta 50% igualitario obteniendo 6.882 registros para cada uno (generando una duplicidad en registros de 1 para balancear). Con la metodologia Smote logramos tener un balanceo de la base 80% para la categoría 0 con 5.444 y 20% para la categoría 1 con 1.361, reduciendo datos de respuesta cero para generar dicho balanceo. Ambas metodologías se van a revisar para tomar la mejor decisión.

Buscamos omitir de la base de datos principal la variable id, la cual representa una identificación del cliente la cual no será necesaria tener presente al momento de ejecutar la modelación. Además, se omiten las variables Fecha de nacimiento y Fecha inicio contrato después de realizar la siguiente transformación a mencionar:

Se genera una construcción de la variable días_de_contrato y edad mediante una ingeniería de variables obtenida mediante el uso de las variables tipo fecha (“Fecha inicio de contrato”) y (“Fecha de nacimiento”). La primera transformación consta en poder obtener los días de contrato que han transcurrido desde que se ejecutó la firma, tomando como comparativo de fecha base el primer día del año 2019. La segunda transformación consta en obtener la edad de la persona tomando como base su fecha de nacimiento, cabe resaltar que la edad obtenida llega hasta la fecha base que se ha planteado, es decir, edad actual en un periodo de tiempo del año 2019. Lo anterior es realizado con el fin de poder medir el impacto de las variables tipo fecha, de manera que podamos tener una lecturabilidad de información más

adecuada y tengamos, así mismo, un uso de variables completas evitando perder información.

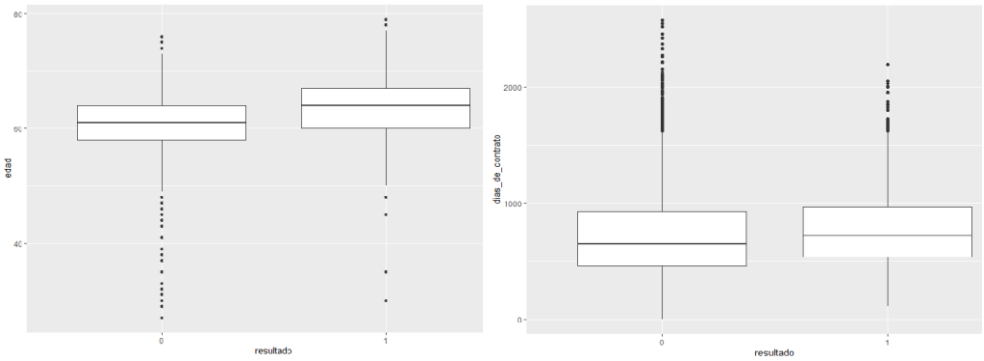

Figura 5. Boxplot edad y días de contrato diferenciado por la deserción

Se procede a convertir tipo factor para poder convertir en dummy la variable tipo cliente en donde se omitirá la variable categórica creada tipo cliente 3 (Variable con menor cantidad de 1 asociada a Cliente empresarial) para evitar multicolinealidad y se dejan las variables tipo de cliente 1: Hombre, y tipo de cliente 2: Mujer.

La variable factura online y plan de datos deben ser tomadas como un factor, sin embargo, su respuesta ya está segmentada a un resultado de 0 y 1, por lo tanto, no es necesario dumificarlas.

4. MODELACIÓN

Selección de Técnicas de Modelación para reducir el AUC.

Para el proceso de modelado se utilizó varios algoritmos como fue regresión Logistica, Redes Neuronales, AdaBoost, XGBoost, LGBMClassifier y CatBoost, este último fue el que se seleccionó al final dado que entregaba mejores resultados con el mismo conjunto de datos.

CatBoost es un algoritmo para potenciar gradientes en arboles de decisión, durante su proceso de entrenamiento se construye un conjunto de árboles de decisión de forma consecutiva y a partir de cierta cantidad de iteraciones se van mejorando los árboles con pérdidas reducidas a comparación de sus antecesores.

Para la generación del mejor modelo se hizo una partición del conjunto de datos en 5 secciones iguales y se hizo un proceso de cross-validation, también se seleccionó dentro de los parámetros del algoritmo aquellos que tenían la capacidad de mejorar el modelo, luego se creó una grilla en donde cada parámetro contenía un conjunto de valores que iban a ser evaluados, en la tabla inicial se observa los rangos que se manejaron inicialmente y los que se obtuvieron después de varias iteraciones del algoritmo.

Tabla 4 Acotación de parámetros en el algoritmo

	learning_rate	**iterations**	**l2_leaf_reg**	**depth**

Rango Inicial	0.01 - 0.1	400 - 2000	0.0 - 3.0	2 - 11
Rango Final	0.03 - 0.05	900 - 1500	0.5 - 1.0	2 - 4

learning_rate: Ayuda a regular la velocidad de en la cual el algoritmo aprende, pequeños valores pueden causar un bajo entrenamiento del modelo si no se tiene la cantidad de iteraciones necesarias, por el contrario, un valor alto puede llegar a causar overfitting.

Iterations: Este valor está muy relacionado con la tasa de aprendizaje por lo que puede causar los mismos inconvenientes en el modelo.

l2_leaf_reg: Coeficiente de regularización L2 para la función de costo

Depth: Profundidad del árbol, al tener valores pequeños de este parámetro se mejora el accuracy de los modelos, pero se requiere una óptima relación entre la cantidad de iteraciones y la tasa de aprendizaje.

Por otro lado, también se realizó el entrenamiento guardando la misma proporción de datos de la variable objetivo *(resultado),* y los datos balanceados por Oversampling y SMOTE con las proporciones mencionadas anteriormente.

Para el caso de los datos originales se obtuvo los resultados de la Tabla 5, donde se observa los valores tomados por los parámetros alcanzado accuracy del 0.885 para el mejor modelo.

Tabla 5 Mejores modelos (balanceo original)

depth	learning_rate	l2_leaf_reg	rsm	border_count	iterations	Accuracy
2	0.05	1.0	0.95	64	900	0.8853575
2	0.04	0.5	0.95	64	1500	0.8851147
2	0.05	1.0	0.95	64	1300	0.8849931

Para el conjunto de datos balanceados por Oversampling se obtuvo los siguientes resultados, se obtienen valores muy cercanos al modelo pasado.

Tabla 6 Mejores modelos (Oversampling)

depth	learning_rate	l2_leaf_reg	rsm	border_count	iterations	Accuracy
3	0.05	1.0	0.95	64	1500	0.8498974
3	0.05	0.5	0.95	64	1500	0.8488802
3	0.05	0.5	0.95	64	1300	0.8432854

Por último, se utilizó los datos balanceados por SMOTE, para este caso se obtuvo peores métricas a comparación de los 2 modelos previos.

Tabla 7 Mejores modelos (SMOTE)

depth	learning_rate	l2_leaf_reg	rsm	border_count	iterations	Accuracy
3	0.05	0.5	0.95	64	1500	0.7857691
3	0.05	1.0	0.95	64	1500	0.7851383
3	0.05	0.5	0.95	64	1300	0.7780018

Para los diferentes modelos se obtuvo la siguiente importancia de las variables predictoras, se puede observar que la variable mora, días de contrato, facturación y edad son predominantes en los modelos que tuvieron mejores métricas.

Original

custom variable importance

	Overall
mora	31.04327
dias_de_contrato	23.56955
facturación	15.91834
edad	14.08241
tipo_cliente_1	5.39361
minutos	4.10927
tipo_cliente_2	2.92879
Antigüedad_Equipo	2.81327
Plan_de_datos	0.12291
Factura_online	0.01856

Oversampling

custom variable importance

	Overall
mora	26.5062
dias_de_contrato	20.8753
facturación	15.3808
edad	14.3550
minutos	7.8369
Antigüedad_Equipo	7.0850
tipo_cliente_1	4.6261
tipo_cliente_2	2.5945
Plan_de_datos	0.4307
Factura_online	0.3095

SMOTE

custom variable importance

	Overall
Antigüedad_Equipo	38.6791
mora	20.0279
dias_de_contrato	13.3678
facturación	9.3465
edad	9.2573
minutos	3.8646
tipo_cliente_1	3.2395
tipo_cliente_2	1.9209
Factura_online	0.1867
Plan_de_datos	0.1097

Figura 6. Importancia de variables por cada modelo

Cada uno de los 3 modelos fue testeado en Kaggle donde se obtuvo los siguientes resultados en términos de AUC

Tabla 8 Validación en Kaggle

Modelo	AUC
Modelo (Balanceo original)	0.89193
Modelo (Oversampling)	0.89202
Modelo (SMOTE)	0.88749

Técnica de modelación para optimizar el kpi de ganancia neta

Conforme a poder interpretar de una manera más adecuada la posible optimización del kpi de ganancia neta, y basándonos en la obtención de coeficientes para lograr tener la matriz de confusión y la curva ROC, nos basamos en generar un modelo stepwise segmentando la base de train, en un aprendizaje 70%-30%. Con lo anterior, y el modelo anteriormente mencionado, obtenemos los siguientes resultados que se ilustran en la Figura 7, por un lado, se destaca que los atributos de minutos y plan_datos0 no tienen influencia en el modelo, y factura_online0 si cuenta con un poco de influencia más no tan significativa como la de los demás atributos.

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.404e+01  6.143e-01 -22.852   <2e-16 ***
factura_online0 -1.546e-01  8.337e-02  -1.855   0.0637 .
plan_datos0     -1.684e-01  1.147e-01  -1.468   0.1422
facturacion      8.562e-06  4.651e-07  18.411   <2e-16 ***
mora             2.237e-02  2.524e-03   8.863   <2e-16 ***
minutos          6.126e-05  3.846e-05   1.593   0.1112
dias_de_contrato 1.452e-03  1.484e-04   9.786   <2e-16 ***
edad             1.316e-02  7.259e-04  18.134   <2e-16 ***
tipo_cliente_1  -1.701e+00  1.135e-01 -14.984   <2e-16 ***
tipo_cliente_2  -1.248e+00  1.096e-01 -11.389   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5563.2  on 6181  degrees of freedom
Residual deviance: 4562.3  on 6172  degrees of freedom
AIC: 4582.3
```

Figura 7. Coeficientes resultado modelo para la optimización Kpi

Tabla de Odds:

Tabla 9 Odds modelo para la optimización Kpi

(Intercept)	factura_online0	plan_datos0	facturacion	mora
8.013422e-07	8.567379e-01	8.450393e-01	1.000009e+00	1.022621e+00
minutos	dias_de_contrato	edad	tipo_cliente_1	tipo_cliente_2
1.000061e+00	1.001453e+00	1.013251e+00	1.825478e-01	2.870254e-01

Adicional a los coeficientes del Figura 8, nos basamos en empezar a interpretar la matriz de confusión y las métricas asociadas. A continuación, se evidencian los resultados:

```
          Reference
Prediction    0    1
         0 5055  802
         1   99  226
> mconftrain$byClass
          Sensitivity          Specificity       Pos Pred Value
            0.21984436           0.98079162           0.69538462
       Neg Pred Value            Precision               Recall
            0.86306983           0.69538462           0.21984436
                   F1            Prevalence       Detection Rate
            0.33407243           0.16628923           0.03655775
Detection Prevalence    Balanced Accuracy
            0.05257198           0.60031799
```

Figura 8. Métricas asociadas a la matriz de confusión

Según lo anterior, el modelo está identificando 226 verdaderos positivos, 5.055 verdaderos negativos, 99 falsos positivos y 802 falsos negativos. Con este primer acercamiento, notamos que el churn rate no está siento alta, además, será importante reducir la tasa de error de los falsos positivos.

Análisis Curva ROC

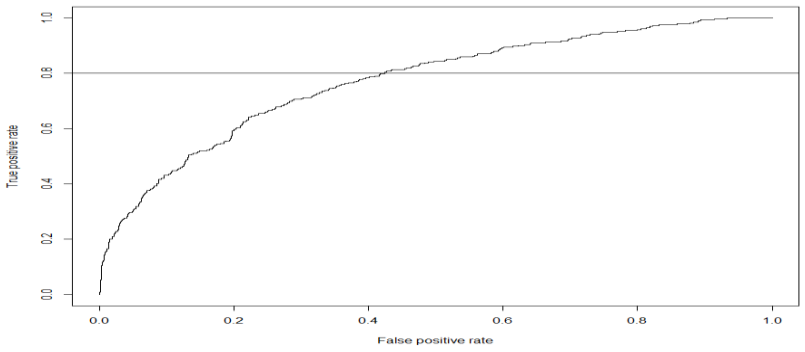

Figura 9. Curva ROC modelo optimización Kpi

Para obtener un acierto del 80% de verdaderos positivos, debemos incurrir en una tasa de 41% de falsos positivos. Sin embargo, el número total de positivos es bajo, por lo cual el porcentaje de 41% no se torna representativo. Por lo tanto, es válido renunciar a tener una precisión alta para poder mejorar la sensibilidad del modelo y tener una mayor proporción de casos positivos, ya que es vital para el negocio encontrar una mayor cantidad de personas que se quieren retirar del plan empresarial vs los que no.

5. EVALUACIÓN

Una vez realizada la implementación de los diferentes modelos, podemos obtener una respuesta con respecto a los resultados más adecuados obtenidos. Procedemos a evaluar los objetivos de negocio:

1. KPI Ganancia Neta: Según las métricas obtenidas anteriormente, y basándonos en el informe que la empresa nos dio respecto a costo/beneficio planteamos la siguiente función para interpretar si se está generando una ganancia con el modelo stepwise utilizado para este proceso.

$$KPI\ GN = B * Recall(\$Ahorro * \frac{\$Costo}{Precisión})$$

$$KPI\ GN = 0{,}1651 * 0{,}38596459(5 - \frac{1}{0{,}6666667})$$

$$KPI\ GN = 0{,}2230$$

Con lo anterior, podemos cumplir con el objetivo de tener una ganancia basándonos en una estrategia de retención de usuarios, mas no de conseguir nuevos clientes, ya que el mercado se encuentra sobre saturado y los usuarios no están de forma constante dispuestos a iniciar un nuevo proceso. Además, le recomendamos a la empresa Claro que su estrategia se base en aumentar la sensibilidad, es decir la tasa de verdaderos positivos y reducir la precisión (Ambas variables tienen una relación inversa). Es importante tener una predicción alta en los falsos positivos, ya que una estrategia de retención impacta los ingresos.

Aunque el mercado está creciendo a un ritmo constante y elevado 11%, el costo de obtener un cliente nuevo es cinco veces superior al de retenerlo, dado a que Claro tiene casi el 50% de los clientes y presenta una ventaja significativa contra sus competidores en términos de esta cuota de mercado, y dado que este crecimiento no ha significado un aumento en los ingresos del sector sino en cambio una disminución, es importante que la estrategia se focalice en la retención de usuarios ya que esto generará eficiencia en costos e impactara positivamente en los resultados de la empresa.

Las estrategias que le recomendamos a Claro constan en: Enfocarse en retener clientes con facturas de alto costo, ya que tienen una propensión mayor a salir, mediante una estrategia de replanteamiento de sus servicios en vía a reducir estos altos costos y generar un servicio optimo y personalizado para los usuarios, evitando gastos innecesarios en productos de poco valor. Además, vemos favorable tener una estrategia de mora, principalmente para evitar que las personas caigan en esta variable mediante un incentivo de reducción de 10% del total de la factura si se cancela el servicio en los primeros 5 días hábiles u incentivos externos para acceder a beneficios adicionales como rifas, descuentos con aliados y promociones. Por último, vemos factible generar beneficios por antigüedad, en donde las personas con mayor tiempo con Claro podrán acceder a descuentos en plataformas de streaming, cupones válidos para redimir en aliados (Cines, supermercados) o acceder a paquetes de productos premium de menor costos que premian la fidelidad.

2. Para predecir la deserción (churn) se puede utilizar algoritmos de boosting que tienden dar mejores métricas para predecir variables categóricas, por otro lado, conservar las proporciones originales y hacer oversampling para obtener la misma cantidad de muestras de ambas categorías tienen comportamiento y métricas similares en el entrenamiento y validación de los datos, en cambio, el modelo con SMOTE tuvo un peor desempeño y esto se puede dar por la pérdida de información que conlleva el algoritmo dado que quita muestras de la categoría con mayor cantidad de observaciones hasta llegar a un porcentaje 80%-20% que fue el que se seleccionó.

Para el caso de los parámetros de los algoritmos que usan arboles de decisión se llega a la conclusión que tener tasas de aprendizaje bajas y profundidades en los árboles pequeños se pueden obtener buenos modelos que no se sobre ajusten y con buenos rendimientos.

Como contraparte a este tipo de algoritmos es que son poco explicativos por lo tanto es difícil entender cómo puede afectar cada variable a la probabilidad de que haya deserción, cosa que resulta muy útil al momento de generar estrategias que impacten el negocio.

BIBLIOGRAFÍA

"Overview - CatBoostClassifier | CatBoost." https://catboost.ai/en/docs/concepts/python-reference_catboostclassifier (accessed May 02, 2022).

"Abonados, ingresos y tráfico de telefonía móvil | Postdata." https://postdata.gov.co/dataset/abonados-ingresos-y-tr%C3%A1fico-de-telefon%C3%ADa-m%C3%B3vil (accessed May 02, 2022).

Chiu, Y. W. (2016). *R for Data Science Cookbook*. Van Haren Publishing.