Universidad del Desarrollo
Facultad de Ingeniería

PROSPECCIÓN DE FUGA DE CLIENTES DE UN SERVICIO DE SUSCRIPCIÓN DIGITAL.

POR: CARLOS ELÍAS PÉREZ PIZARRO

Proyecto de grado presentado a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science

PROFESOR GUÍA:
Dra. LORETO BRAVO CELEDÓN.

Enero 2022
SANTIAGO

Resumen

En el mundo actual donde la incertidumbre pareciera estar en todos los aspectos de nuestra vida, sumado al creciente acceso a internet por parte de las personas, es donde los periódicos han visto la oportunidad de ampliar su alcance y ser capaces recuperar las audiencias perdidas, debido a la lenta pero inevitable migración desde el tradicional periódico impreso hacia las plataformas digitales. Con el fin de crear una relación más directa con sus audiencias, que a su vez permita proyectar con mayor precisión los ingresos futuros, es que los medios escritos de prensa alrededor del mundo han implementado modelos de suscripción digital, los que, gracias al avance en las tecnologías de la información, han posibilitado la acumulación de grandes volúmenes de datos, sobre el uso que realizan los usuarios del servicio. Lo anterior nos lleva de forma natural a preguntarnos cuáles son las características que comparten estas personas, desde el momento en que se suscriben, hasta que, en algunos casos, deciden dar de baja el servicio.

Este trabajo utiliza los datos contenidos en el contrato de cada cliente del servicio de suscripción digital de La Tercera, lanzado en abril de 2019, sobre los cuales se aplicarán técnicas de machine learning y algoritmos como K-Means, para segmentar a los suscriptores de acuerdo con la frecuencia de uso del servicio. Posteriormente se evidenciará mediante el modelo de supervivencia propuesto por Kaplan-Meier y el modelo de riesgo proporcional (Cox), que un suscriptor que navega diariamente tiene un 72% menos de riesgo de fuga comparativamente a uno que lo realiza de forma esporádica. Utilizando el mismo método se mostrará la diferencia en riesgo de fuga que existen entre la elección de un plan digital (+60% riesgo) por sobre uno con producto impreso, entre otras variables analizadas. De esta manera, se podrán tomar acciones comerciales preventivas, que apunten disminuir la cantidad de usuarios fugados tanto de forma voluntaria como involuntaria.

Palabras clave: Análisis de supervivencia, Churn, Suscripción, Cox, Kaplan-Meier.

1. Introducción

Para nadie es desconocido que a nivel mundial la prensa escrita ha experimentado durante años una caída sostenida en la circulación impresa de sus ediciones, debido al aumento en el acceso a la información propiciado por la masificación de internet, el cual ha sido amplificado por su fusión con la telefonía portátil, haciendo posible que las personas se encuentran conectadas e informadas las 24 horas del día. Como respuesta natural a este fenómeno, es que medios mundialmente conocidos como *The New York Times* o *The Washington Post*, han revolucionado la forma de entregar contenido informativo, con sus servicios de suscripción digital, lo que les ha permitido a los usuarios tener acceso a información de calidad donde quiera que estén, en una amplia variedad de plataformas, permitiendo a su vez a las editoriales, migrar desde un modelo de ingresos basado en publicidad, a uno en donde los suscriptores son la base del negocio [11].

De forma local en Chile, el mercado se encuentra dominado por dos conglomerados periodísticos, *Grupo Mercurio* con amplia cobertura nacional gracias a su red de diarios regionales y *Grupo Copesa*, quien ha apostado en la transformación digital de sus plataformas, con la creación de un servicio de suscripción digital para *La Tercera*, su medio más destacado, la que desde su creación en el segundo trimestre del 2019, ha reportado un sostenido aumento en la cantidad de suscriptores abonados a sus distintos tipos de planes.

Como todo producto o servicio introducido al mercado, un servicio de suscripción suele presentar una curva de crecimiento con pendiente positiva, dado que se encuentra en las primeras etapas de su ciclo de vida [13], pero a medida que esta curva comienza a aplanarse, cobra mayor importancia la capacidad de tener una fuga controlada, que permita mantener un parque de suscriptores estable, aun cuando la venta pueda sufrir altibajos, ya sea por la estacionalidad propia del mercado o por encontrarse más cercana a la maduración en el ciclo de vida del producto.

Cuando hablamos de fuga, la industria y la literatura reconoce dos grandes grupos [4], la fuga voluntaria que se explica por factores como no estar de acuerdo con la línea editorial del medio, problemas de login de la cuenta, aumentos de precio u otras que dependan de una decisión activa del cliente y por otro lado, la fuga involuntaria, la que ocurre cuando el cliente deja de pagar su suscripción, debido a problemas al momento de realizar el cargo en su tarjeta. En este trabajo analizaremos las variables que afectan a ambos tipos de fuga.

Una forma de abordar la fuga es a partir del ciclo de vida del cliente, en donde nuestro objetivo es predecir el momento en el futuro en que este dejará de formar parte del parque de suscriptores, para poder tomar acciones comerciales preventivas antes de que esto suceda. Lo anterior es importante, ya que es más eficiente económicamente mantener un cliente activo que obtener uno nuevo [10].

Este trabajo implementará un algoritmo de machine learning basado en el método no determinístico de Kaplan-Meier, el cual entrega la probabilidad de supervivencia luego de ocurrido un determinado evento [1] de acuerdo con el modelo de riesgo proporcional de Cox [6], el cual indica la posibilidad de ocurrencia del evento usado en Kaplan-Meier.

2. Trabajo relacionado

El uso de análisis de supervivencia se encuentra extensamente documentado desde 1958, año en el que Edward L. Kaplan y Paul Meier publicaron el estudio *Nonparametric Estimation From Incomplete Observations*, en donde señalan que, tanto en el ámbito médico como en otros campos, la ocurrencia de "la muerte" se puede prevenir para algunos miembros de la muestra, por la ocurrencia previa de algún otro evento [1]. Este "otro evento" pasó a formalizarse como aquellas covariables que afectan la vida de cada individuo que forma parte del análisis de supervivencia y se denominó Regresión de Cox o Modelo de riesgos proporcionales [6].

A medida que pasaron los años, con el avance de los distintos algoritmos de clasificación basados en técnicas de machine learning como Random Forest, AdaBoost o Regresión Logística entre otros, se hizo posible determinar con un nivel de exactitud superior al 85% [2] el estado activo o fugado en distintos tipos de bases de datos, siendo uno de los más populares y relevantes para este estudio, dada la similitud al ser ambos modelos de suscripción, el análisis de propensión de fuga de clientes de empresas telefónicas [3]. Sin embargo, este tipo de algoritmos no entregan información crucial para una compañía, como lo es el tiempo que falta para que ocurra la baja del plan, algo que el análisis de supervivencia no solo proporciona, sino que complementa con la probabilidad que tiene el suscriptor en cada momento en el tiempo de sobrevivir o no [3].

Como era de suponer, dado los beneficios del análisis de supervivencia por sobre otras algoritmos de machine learning, este empezó a ser incorporado como librería en los lenguajes más usados en la ciencia de datos, como lo son Python y R. Su forma más conocida es la librería Lifelines, la que se encuentra ampliamente documentada [4], aunque otras librerías como scikit-learn también lo incluyen [5].

Cuando nos enfrentamos a grandes cantidades de datos, la técnica de segmentación K-Means [7] suele ser usada para encontrar patrones subyacentes en ellos, que ayuden a vislumbrar comportamientos que a simple vista pueden no ser tan evidentes. Este algoritmo suele ocuparse junto con el análisis denominado RFM *(recency, frequency and monetary)* [8], para determinar el nivel de *engagement* que tienen los miembros del estudio con el producto o empresa. En este estudio se utilizará solo la variable F "*frequency*", para determinar la frecuencia de uso que tienen los suscriptores, respecto de su navegación en www.latercera.com durante los 30 días previos a su fuga.

3. Hipótesis y Objetivos

3.1. Hipótesis

La implementación de un modelo de machine learning basado en análisis de supervivencia de Kaplan-Meier y el modelo de riesgo proporcional (Cox), modelará las variables que afectan a la fuga de suscriptores del servicio de suscripción digital de La Tercera.

3.2. Objetivos

- **General**: Implementar un algoritmo de machine learning basado en análisis de supervivencia de Kaplan-Meier y el modelo de riesgo proporcional (Cox), que permitan identificar la probabilidad de fuga que tiene los suscriptores, de acuerdo con las características propias de su contrato y la intensidad de uso de la suscripción digital de La Tercera.

- **Específicos:**

1. Determinar mediante K-means, los distintos grupos de usuarios de la suscripción digital de La Tercera, basado en su frecuencia de navegación.

2. Obtener mediante análisis de supervivencia y el modelo de riesgo proporcional (Cox), las covariables que afectan el tiempo de permanencia de una persona como suscriptor de La Tercera.

4. Datos y Metodología

4.1. Datos

La base de datos a consultar es de propiedad del Consorcio Periodístico de Chile (COPESA S.A).

Datos de suscriptores: Esta base contiene el registro histórico de cada contrato que ha sido vendido por La Tercera desde el año 2014 a la fecha. Contiene 119 columnas y más de 500.000 filas (figura 1), con información relativa al contrato de venta de la suscripción, como, por ejemplo: estado del contrato, fecha de inicio, medio de pago, valor mensual del plan, tipo de plan, motivo de fuga, fecha de fuga, etc., tanto de contratos que actualmente se encuentran activos, como aquellos que se han fugado. Para fines de este estudio, se utilizarán los datos de personas naturales, con fecha de inicio y término de contrato, desde mayo de 2019 a noviembre de 2021.

	cliente	organizacion	isactive	co_salesorders_id	folio_contrato	estado_folio	doctype	fecha_estado
0	COPESA S.A.	COPESA S.A.	Y	2695636	1627074	CERRADO	POS Order	2020-01-31 08:56:03
1	COPESA S.A.	COPESA S.A.	Y	2695360	1627074	CERRADO	POS Order	2020-01-31 08:56:03
2	COPESA S.A.	COPESA S.A.	Y	2695637	1627074	CERRADO	POS Order	2020-01-31 08:56:03
3	COPESA S.A.	COPESA S.A.	Y	3413128	1627074	CERRADO	POS Order	2020-01-31 08:56:03
4	COPESA S.A.	COPESA S.A.	Y	3413129	1627074	CERRADO	POS Order	2020-01-31 08:56:03

Figura 1: Referencia de dataframe de base de datos de clientes.

4.2. Descripción de los datos

El base de datos depurada contiene 39.339 registros los cuales se encuentran divididos en 27.596 suscriptores activos (70.1%) y 11.743 fugados (29.9%) como se muestra en la figura 2.

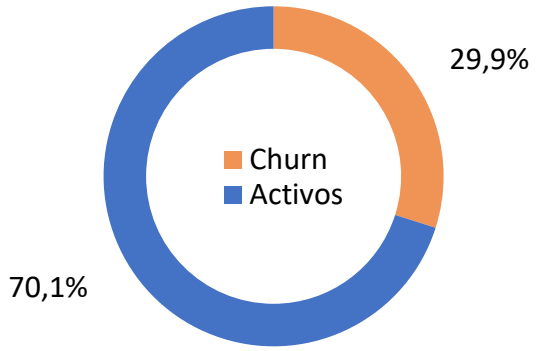

Figura 2: Proporción de clientes activos y fugados considerados en el estudio.

A continuación, se revisarán las principales distribuciones de aquellas variables continuas que se utilizaron para construir el modelo predictivo, diferenciando aquellos contratos que se encuentran en estado activo codificado como “0” y fugado con etiqueta “1”.

a. Valor mensual promedio: Este variable fue construida a partir del total de los pagos que ha realizado el cliente desde el inicio de su suscripción, incluyendo el valor promocional que se entrega al inicio de la contratación del plan, dividido en la antigüedad del contrato expresada en meses.

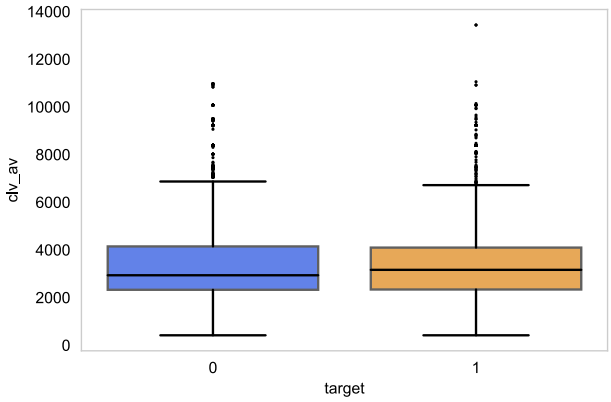

Figura 3: Boxplot de valor mensual promedio según estado del contrato.

Como podemos ver en la figura 3, para ambos conjuntos de datos la mediana es ligeramente superior $3.147 (+9.5%) en los contratos fugados, comparada con aquellos activos $2.873.

b. Antigüedad del cliente: Esta fue calculada como la diferencia entre la fecha de inicio de la suscripción y la fecha de término del contrato para el caso de los suscriptores fugados. Para los folios activos, se calculó la diferencia entre la fecha de inicio y la fecha de corte del estudio 30/11/2021. Ambas fueron expresadas en meses.

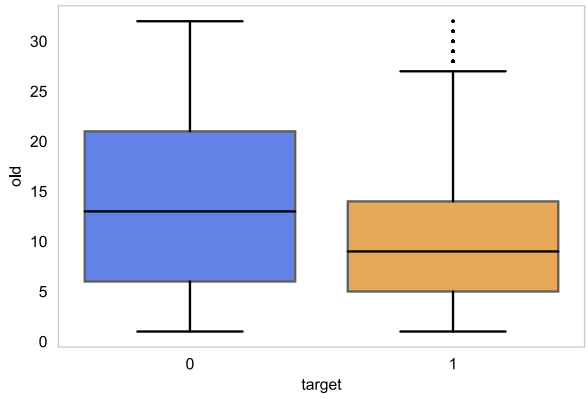

Figura 4: Boxplot de antigüedad de suscriptores representada en meses.

Se aprecia en la figura 4 que para este conjunto existe mayor dispersión de los datos entre aquellos que se encuentran activos, quienes presentan un rango intercuartil entre 6 y 20 meses con una mediana de 13, mientras que aquellos fugados varían entre los 5 y 13 meses, con una mediana de 9.

c. Porcentaje de descuento promocional: Corresponde a la variación entre el precio pagado por el cliente al inicio de su contrato y el valor una vez terminado el periodo promocional.

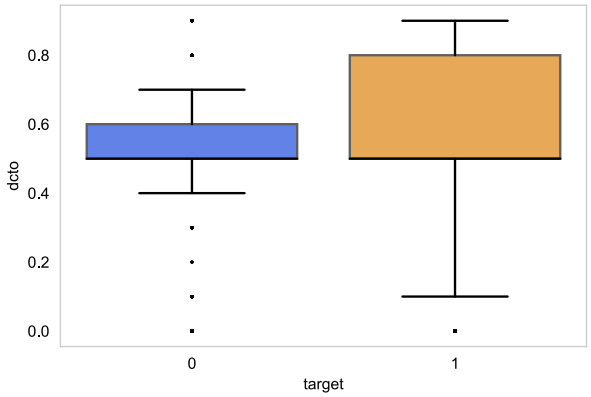

Figura 5: Boxplot del porcentaje de descuento promocional.

La figura 5 presenta una diferencia entre los clientes activos y fugados, en donde, a pesar de tener una mediana de 50% de descuento en ambos casos, la distribución es distinta en ambos conjuntos, llegando hasta 80% descuento en su tercer cuartil para los contratos fugados.

d. Frecuencia de navegación: Corresponde a la cantidad de visitas únicas realizadas por el suscriptor al sitio www.latercera.com en los últimos 30 días previos al cierre de su contrato o al cierre del estudio para los casos activos.

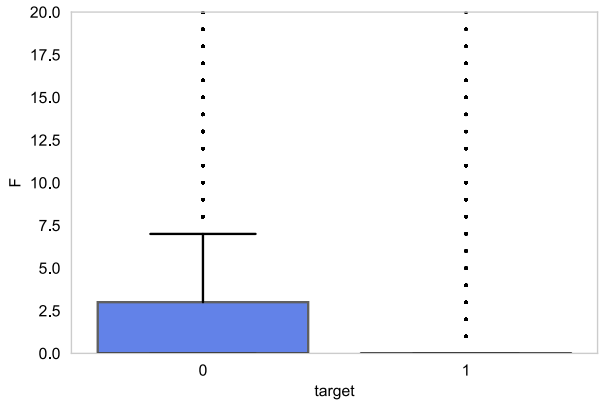

Figura 6: Boxplot de la frecuencia de navegación últimos 30 días.

La figura 6 evidencia como los suscriptores activos presentan una mayor frecuencia de uso de la suscripción digital, aun cuando ambos comparten una mediana en cero, la distribución de aquellos activos alcanza hasta 3 visitas mensuales en su tercer cuartil.

4.3. Metodología

La metodología fue diseñada para trabajar sobre datos de los suscriptores de La Tercera obtenidos de la base ya mencionada de acuerdo con los siguientes pasos.

1. Consulta de los datos: Esta base de datos se encuentra disponible mediante archivo en formato csv, generado diariamente a través de proceso automático del departamento de Tecnología. La base incluye movimientos como la creación de nuevos contratos, procesos de cierre por fuga, upselling, etc, generados hasta el día anterior.

2. Preprocesamiento:

- Se construye pipeline para la obtención de las columnas relativas al problema, dejándolas en el formato que corresponde, por ejemplo, columnas con datos de fecha, pasan desde formato string a date.

- Se enmascaró la identidad de los suscriptores usando como llave su email, el cual fue encriptado utilizando el método MD5 de la librería hashlib.

- Se realizan filtros necesarios para obtener los datos desde mayo 2019 a noviembre de 2021, de personas naturales, con contratos en estado activo y cerrado, que tengan acceso digital, con un valor neto mayor a $800 mensuales, con pago mensual recurrente, de los canales de venta internet y call center.

- Feature engineering: Se crean nuevos datos a partir de los existentes:

    - Valor Promocional (promo): Valor 0 ó 1 para determinar si el contrato tuvo o no valor promocional de entrada.
    - Descuento de entrada (dcto): Rango de valores decimales entre 0.0 y 0.8 para determinar el porcentaje de descuento promocional del plan.
    - Antigüedad (old): Numero entero expresado en meses para determinar el tiempo transcurrido entre el inicio del plan, el presente o el fin del contrato.
    - Precio promedio (clv_av): Numero decimal, expresado en pesos chilenos, para determinar el valor cuota promedio que paga cada suscriptor.
    - Frecuencia (F): Numero entero para determinar la frecuencia de uso de la suscripción digital.
    - Cluster de frecuencia (F_Cluster): Valores 0, 1 ó 2 para segmentar la variable F en grupos de usuarios.

- Para el resto de las variables categóricas como “termino_pago”, “superplan” y “canal_venta”, se utilizó la función *get_dummies* de la librería Pandas de Python, con el fin de convertirlas en variables binarias.

- Segmentación: Se realiza segmentación de la navegación de los usuarios, utilizando el algoritmo K-Means de la librería sklearn. En primera instancia se planificó hacer la segmentación utilizando una variante del método RFM, conocida como FRM-I [15], pero al realizar análisis de componentes principales PCA [16] se determinó que el 90% de la variabilidad de los datos proviene de la característica frecuencia de uso (F), por lo que la segmentación fue aplicada solo sobre esta, con el fin de disminuir el tamaño final de la base y por ende disminuyendo el tiempo de procesamiento. Para el caso de aquellos suscriptores fugados, se consideran los 30 días previos al cierre de su contrato, mientras que, para los suscriptores activos, se considera 30 días hasta la fecha de cierre de este estudio (30/11/2021 inclusive).

  Para determinar la cantidad de segmentos a obtener, se utilizó la métrica “método del codo” la cual entrega la cantidad optima de grupos que mejor representa una población [9]

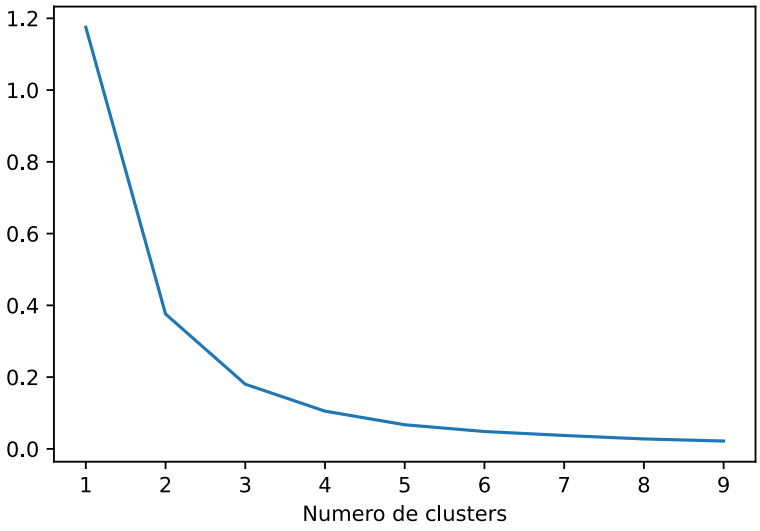

Figura 7: Método del codo aplicado sobre navegación de suscriptores.

La figura 7 muestra que el "método del codo" sugiere que 3 es el número óptimo de segmentos con los que se debe construir esta agrupación.

- Análisis de supervivencia de Kaplan-Meier: Es un método no paramétrico [14], también conocido como Estimador del producto-límite [3]. En este método el estimador $\hat{S}(t)$ (figura 8), es la probabilidad de sobrevivir más allá de t, se calcula para cada t, en donde:

    - t es el mes en el que ocurre la fuga.
    - d_i es el número de contratos cerrados en el momento t_i.
    - n_i es el número de suscriptores en riesgo justo antes de t_i.

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}$$

Figura 8: Estimador de supervivencia de Kaplan-Meier

- Modelo de riesgos proporcionales de Cox (CPH): El modelo CPH determina el efecto que tendrá el cambio de una unidad de una covariable sobre la probabilidad de supervivencia de una observación [6]. CPH es un modelo semiparamétrico que consta de dos partes: La función de riesgo base y la relación de las covariables con la función de riesgo base.

5. Resultados

5.1. Análisis de supervivencia de Kaplan-Meier

Se procedió a calcular la curva de vida característica de un suscriptor promedio, desde donde se desprende que luego de 20 meses, la suscripción digital de La Tercera tiene una tasa de retención de un 60% (Figura 9).

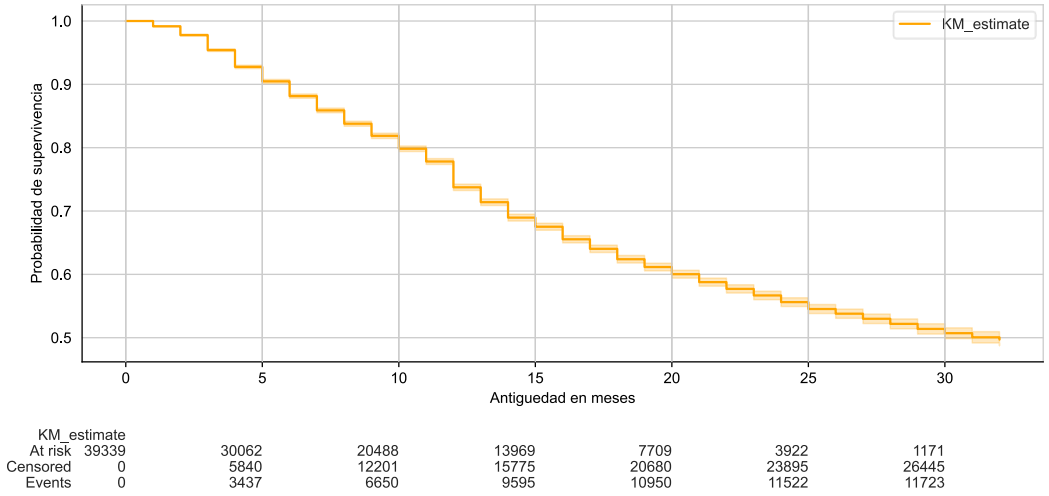

Figura 9: Curva de supervivencia de Kaplan-Meier para todos los suscriptores

A continuación, podemos ver los resultados de las distintas variables propias del contrato de los suscriptores, en donde se compara la probabilidad de fuga (eje Y) en determinado mes (eje X), según análisis de supervivencia basado en el estimador de Kaplan-Meier.

En la figura 10 se observa que los clientes que han contratado un plan digital, tienen una mayor probabilidad de fuga que aquellos que tienen un plan que incluye producto impreso, esto debido, entre otros factores, a la promesa de venta de poder dar de baja el plan digital en cualquier momento.

También se aprecia como luego del primer año de suscripción, los clientes con plan impreso, comienzan a dar de baja sus contratos llegando a igualar la curva de vida de aquellos que son digitales alrededor de los 15 meses (Figura 10). Esto ocurre ya que aquellos clientes que han contratado un plan con producto impreso, tienen trabas al momento de querer dar de baja la suscripción volutariamente, como lo es el cobro de los meses restantes para cumplir la anualidad entre otras.

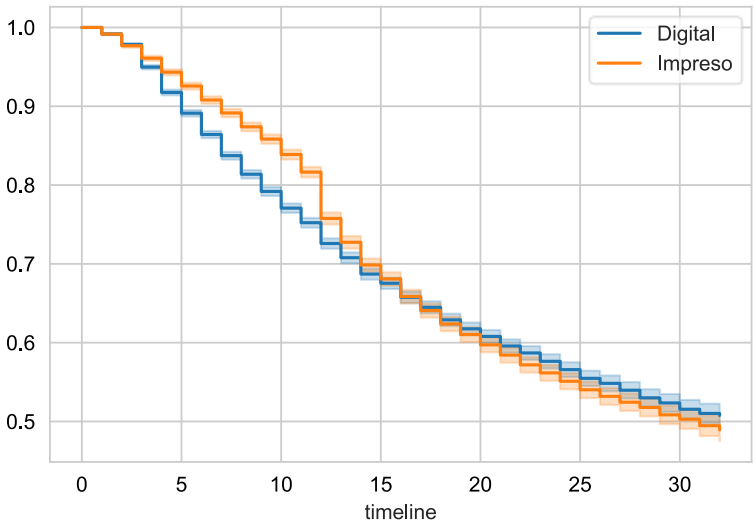

Figura 10: Probabilidad de supervivencia con respecto al tipo de plan.

En la figura 11 se evidencia como el medio de pago PayPal tiene una mayor probabilidad de fuga respecto de MercadoPago, OneClick y PAT, ya que PayPal permite al cliente dar de baja el cobro de forma unilateral, sin intermediar el banco, lo que no ocurre con las otras dos formas de pago. Cabe señalar que MercadoPago tiene una menor cantidad de meses en su curva, ya que la integración de este medio con la plataforma web de venta de suscripciones, se realizó a inicios del año 2020.

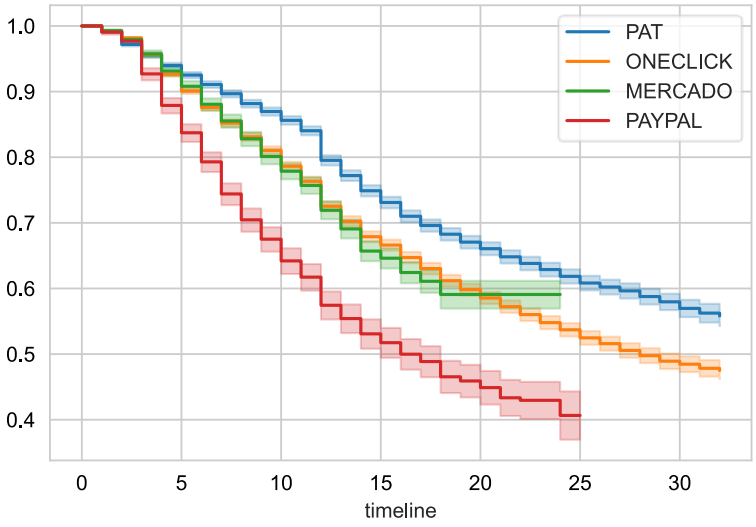

Figura 11: Probabilidad de supervivencia con respecto al medio de pago.

En la figura 12 se observa que a medida que aumenta el descuento de entrada con el que se adquiere la suscripción, es mayor la probabilidad de fuga del cliente, esto se debe entre otros motivos a la gran diferencia que existe, por ejemplo, en planes digitales, entre comenzar pagando $990 mensual (80% descuento por 3 meses), para luego pasar a un valor de $4.990 mensual desde el cuarto mes, lo que equivale a un 404% de aumento. Lo anterior, podría motivar al cliente a dar de baja el servicio al ver reflejado el cobro en su cartola. De forma distinta los planes que pagan $2.994 (40% descuento por 6 meses), al pasar a tarifa completa al séptimo mes, solo tienen un 67% de aumento de precio.

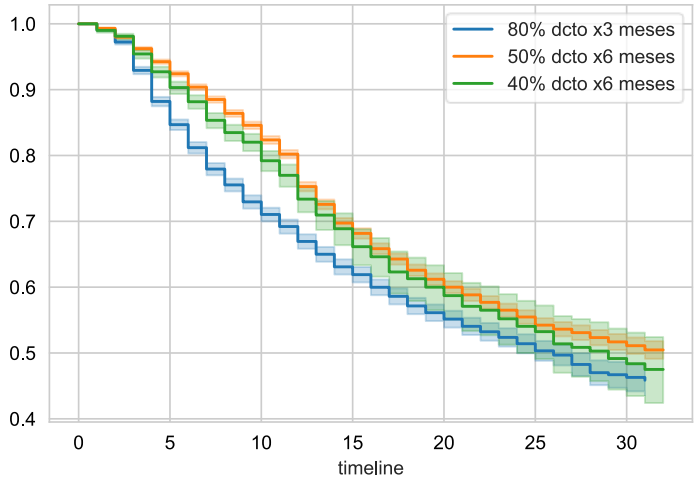

Figura 12: Probabilidad de supervivencia con respecto al descuento de entrada.

Finalmente, la figura 13 muestra el comportamiento de los segmentos obtenidos por el algoritmo K-Means, en donde cada uno de los 3 segmentos (Navegación alta, media y baja), representa la frecuencia de uso del servicio digital. De esto se desprende que entre menos navegación tiene un suscriptor, mayor es su probabilidad de fuga. Cabe señalar que lo amplio de los intervalos de confianza del segmento medio y alto, se debe a que la gran mayoría de los datos están acumulados en el segmento bajo como se aprecia en la figura 14, con un 88% de los casos de clientes activos y un 96% para los fugados.

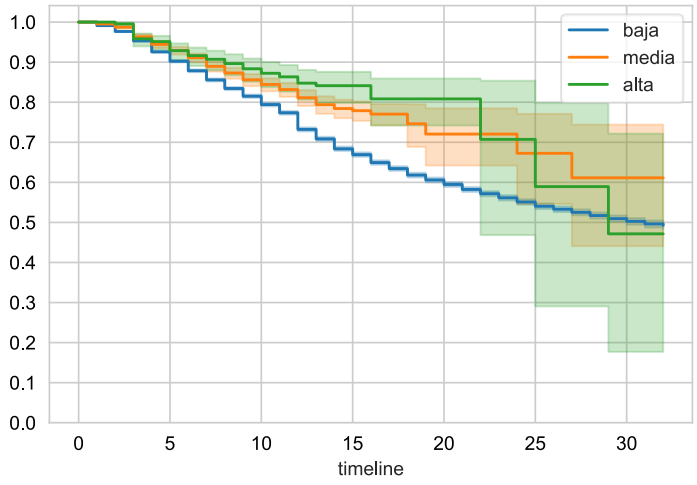

Figura 13: Probabilidad de supervivencia con respecto a segmentación de la variable frecuencia de navegación

target	F_Cluster	count	%	mean	std	min	25%	50%	75%	max
Clientes activos	baja	24.152	88%	1	4	0	0	0	0	19
	media	2.818	10%	38	15	20	26	35	48	74
	alta	626	2%	112	38	75	85	99	126	357
Clientes fugados	baja	11.282	96%	1	3	0	0	0	0	19
	media	387	3%	38	14	20	25	33	48	74
	alta	74	1%	110	38	75	85	100	124	297

Figura 14: Frecuencia de uso de la suscripción digital según target y segmento de navegación.

5.2. Calibración del modelo Modelo CPH

Como parte del proceso de validación, se obtuvo un 86.6% de concordancia, interpretable de forma similar a la curva ROC de una regresión logística, en donde 1.0 representa una perfecta concordancia y 0.5 el resultado esperado de una predicción al azar, de esta forma se puede comprobar si el modelo diferencia entre contratos activos y fugados. Este proceso fue realizado utilizando la función “k_fold_cross_validation” de la librería Lifelines [4]. La figura 15 muestra como para los primeros 4 deciles el modelo está subestimando el riesgo de fuga, mientras que para los últimos 4 está sobrestimado.

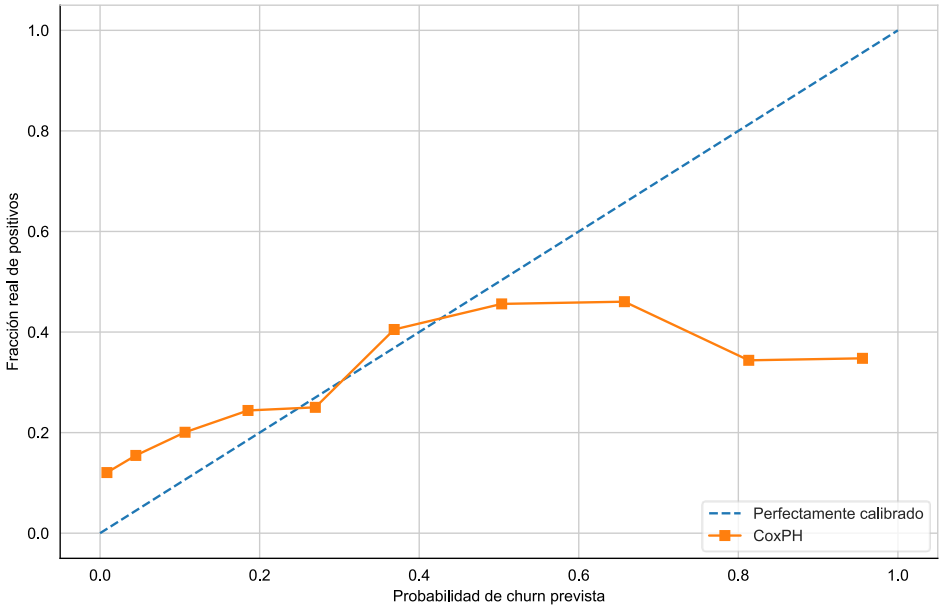

Figura 15: Probabilidad de supervivencia con respecto al descuento de entrada.

De forma paralela se procedió a calcular el score de Brier, función semejante al error cuadrático medio, la cual entregó un 0.15 para 12 meses. Esto fue realizado utilizando la función “calibration_curve” de la librería sklearn [17].

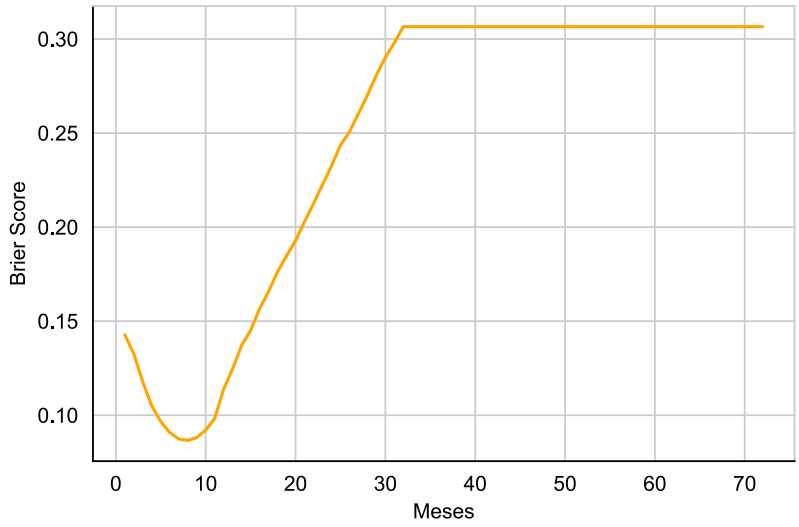

Figura 16: Calibración del modelo de riegos proporcionales de Cox a lo largo del tiempo

De la figura 16 se desprende que para predicciones entre 1 y 12 meses el modelo se encuentra razonablemente calibrado, luego de eso se deteriora rápidamente.

5.3. Análisis de covariables según modelo CPH

Este análisis entrega la interacción que tiene la totalidad de las covariables del estudio con el riesgo base de un suscriptor promedio, mostrando el efecto que tendrá el cambio de una unidad de una covariable sobre la probabilidad de supervivencia de un suscriptor. Las columnas de la figura 17 se explican a continuación:

- **coef:** representa a los coeficientes del modelo, indicando como cada covariable impacta el riesgo de fuga. Un valor positivo indica que un suscriptor tiene mayor probabilidad de fuga, de forma contraria, un valor negativo indica mayor probabilidad de permanencia.

- **exp (coef):** Indica la razón de riesgo de cada covariable, siendo 1.0 el valor neutral.
- **p:** *"p value"*, indica el grado de significancia que tiene esta variable en contraste de la hipótesis nula.

variable	coef	exp(coef)	p
superplan	0.471	1.601	<0.0005
canal_venta	0.321	1.378	<0.0005
dcto	0.279	1.322	0.177
clv	-0.000	1.000	<0.0005
clv_av	-0.001	0.999	<0.0005
p_min	0.001	1.001	<0.0005
p_max	-0.000	1.000	<0.0005
promo	-0.141	0.868	0.020
F	-0.003	0.997	0.186
Nav	-0.339	0.712	<0.0005
F_Cluster	-0.327	0.721	<0.0005
P_Cluster	0.953	2.595	<0.0005
termino_pago_ONECLICK	0.064	1.067	0.031
termino_pago_PAT	-0.044	0.957	0.336
termino_pago_PAYPAL MENSUAL	0.434	1.544	<0.0005

Figura 17: Interacción de covariables en la probabilidad de fuga.

La figura 17 muestra que la variable "superplan" tiene un coeficiente positivo de 0.471, implicando mayor riesgo de fuga para un plan digital, el cual, si se compara con un plan que incluye producto impreso, la elección de digital aumenta en 60% la probabilidad de fuga, según indica el exp (coef) con un p value menor a 0.0005. De la misma forma el medio de pago PayPal aumenta el riesgo de fuga un 54%, mientras que la segmentación representada en "F_Cluster" indica que un cliente que pasa desde un segmento medio de navegación a uno bajo tiene un 72% más de riesgo de fugarse.

6. Conclusiones y trabajo futuro

Como se ha demostrado, es posible modelar los distintos tipos de comportamiento de fuga de acuerdo con las características bajo las cuales son contratados los planes de suscripción a La Tercera. Como principales conclusiones se destaca que la tasa de retención a los 20 meses es del 60% de los clientes, la que depende de variables como el tipo de plan que se contrate, en donde aquellos clientes que eligen un plan digital tendrán un aumento de 60% en su probabilidad de fuga, comparados con aquellos que eligen impreso, dada la promesa de venta que los habilita a dar de baja el servicio cuando quieran. Por otro lado, aquellos planes contratados utilizando el medio de pago PayPal, tendrán un aumento de 54% del riesgo de fuga, debido a la facilidad que tiene el cliente para desinscribir el pago con este medio. Así mismo la frecuencia de uso del servicio es un gran predictor respecto de la cantidad de meses que el cliente será suscriptor, ya que si comparamos un suscriptor que navega en promedio 38 veces al mes, es decir, aproximadamente una vez al día y otro que solo tiene una visita mensual, el riesgo de fuga aumenta un 72%.

Como parte del trabajo futuro, se recomienda incluir variables respecto del uso de la suscripción digital, como tiempo promedio de lectura, secciones visitadas, page views realizadas, cantidad de días desde la última conexión, así como incluir aspectos relativos a la experiencia del consumidor, como la cantidad reclamos realizados en el call center de atención a clientes y la frecuencia de uso del club de fidelización de La Tercera.

Cabe señalar que, al momento de cierre del estudio, se han comenzado a implementar acciones comerciales sobre los suscriptores que tengan una probabilidad de fuga mayor al 70%, como el envío de emails con códigos de cortesía para ser usados en cafeterías y descuentos exclusivos con motivo del cumpleaños del suscriptor, así como también mejoras en la sucursal virtual, relativas al cambio de medio de pago y renovación automática de planes anuales entre otros. Estas acciones deberán ser medidas para cuantificar el impacto que han tenido sobre la curva de vida de los distintos segmentos de clientes.

Referencias

1. Kaplan EL, Meier P (1958) Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association 53:457–481. https://doi.org/10.1080/01621459.1958.10501452
2. Ullah I, Raza B, Malik AK, et al (2019) A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. IEEE Access 7:60134–60149. https://doi.org/10.1109/ACCESS.2019.2914999
3. Masarifoglu M, Hakan Buyuklu A (2019) Applying Survival Analysis to Telecom Churn Data. AJTAS 8:261. https://doi.org/10.11648/j.ajtas.20190806.18
4. Davidson-Pilon C (2019) lifelines: survival analysis in Python. JOSS 4:1317. https://doi.org/10.21105/joss.01317
5. Pölsterl S scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. 6
6. Benítez-Parejo N, Rodríguez del Águila MM, Pérez-Vicente S (2011) Survival analysis and Cox regression. Allergologia et Immunopathologia 39:362–373. https://doi.org/10.1016/j.aller.2011.07.007
7. Likas A, Vlassis N, J. Verbeek J (2003) The global k-means clustering algorithm. Pattern Recognition 36:451–461. https://doi.org/10.1016/S0031-3203(02)00060-2
8. Aleksandrova Y (2018) APPLICATION OF MACHINE LEARNING FOR CHURN PREDICTION BASED ON TRANSACTIONAL DATA (RFM ANALYSIS)
9. Syakur MA, Khotimah BK, Rochman EMS, Satoto BD (2018) Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. IOP Conf Ser: Mater Sci Eng 336:012017. https://doi.org/10.1088/1757-899X/336/1/012017
10. Reinartz WJ, Kumar V (2003) The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration. Journal of Marketing 67:77–99. https://doi.org/10.1509/jmkg.67.1.77.18589
11. O'Sullivan J, Fortunati L, Taipale S, Barnhurst K (2017) Innovators and innovated: Newspapers and the postdigital future beyond the "death of print." The Information Society 33:86–95. https://doi.org/10.1080/01972243.2017.1289488
12. Hadden J, Tiwari A, Roy R, Ruta D (2006) Churn Prediction: Does Technology Matter? International Journal of Electrical and Computer Engineering 7
13. Day GS The Product Life Cycle: Analysis and Applications Issues. 8
14. Wang P, Li Y, Reddy CK (2017) Machine Learning for Survival Analysis: A Survey. arXiv:170804649 [cs, stat]
15. Tkachenko Y (2015) Autonomous CRM Control via CLV Approximation with Deep Reinforcement Learning in Discrete and Continuous Action Space. arXiv:150401840 [cs]
16. Makiewicz A, Ratajczak W PRINCIPAL COMPONENTS ANALYSIS (PCA). Principal Components Analysis 40
17. Pedregosa F, Varoquaux G, Gramfort A, et al Scikit-learn: Machine Learning in Python. MACHINE LEARNING IN PYTHON 6