Master in Management + Analytics

Predicción de Churn de Seguros con LightGBM

Franco Tralice

Director: Ramiro Gálvez - Departamento de Computación, FCEyN, UBA

Abril, 2019

Índice

1- Introducción

1.1 - El Dominio

Ante el riesgo de posibles eventualidades, los individuos buscan disminuir el grado de incertidumbre. Es por eso que buscan adquirir seguros para aumentar su utilidad. En este escenario es donde aparecen compañías aseguradoras que brindan un respaldo ante los clientes que necesitan resguardar sus bienes.

Las compañías de seguro son sociedades que realizan el servicio de cobertura de riesgos, haciéndose cargo parcial o totalmente de los daños económicos que pueden ser causados por accidentes. Las aseguradoras cuentan con ingresos mensuales, que son las cuotas de los clientes que la componen. Los egresos provienen principalmente de los pagos que se hacen a los clientes que tuvieron algún tipo de eventualidad. Es por eso para este tipo de compañías es muy importante mantener una masa crítica tal que sus cuotas puedan hacer frente a todos los egresos mensuales y que la compañía siga en pie. (Guillen, M., Nielsen, J. P., & Pérez-Marín, A. M., 2008).

En el caso particular referido a este trabajo, la aseguradora en cuestión es una empresa que cuenta con muchos clientes a nivel nacional e internacional. Además, está vinculada con un banco importante del país. Es por eso que se pueden aprovechar muchos más datos de los clientes que los que habría en una firma común.

1.2 - El problema

Los esfuerzos de marketing para conseguir clientes nuevos son muy grandes, es por eso que una gran parte de estos mismos consisten en retener a los clientes que ya aportan pagando la cuota de algún seguro a la empresa. Retener un cliente es menos costoso que volver a conseguirlo. Aquí es donde introducimos la métrica que lleva el nombre del trabajo, el *churn* (Soeini, R. A., & Rodpysh, K. V., 2012).

Definimos a la tasa de churn como el *Key Performance Indicator* (KPI) que mide el porcentaje de clientes que dejan de utilizar un producto o un servicio, o dicho más concretamente, la tasa de abandono de clientes. La compañía invierte muchos de sus activos en tratar de bajar lo máximo posible este ratio, que se traduce en más solvencia, más ingresos y por lo tanto más retornos para la misma. Los motivos de las bajas de los clientes son diversos. Entre ellos, podemos mencionar el valor de la cuota, porque no sienten necesario tener un seguro, etc.

En muchos ambientes, y en particular en los de los seguros, la cantidad de clientes es vital para la compañía. Pero, ¿cuánto vale cada uno de los clientes? El *Customer Lifetime Value* (CLV) es el término que se utiliza para la estimación del valor presente neto del fondo de flujos que el cliente brindará a la compañía a lo largo de su vida. La fórmula para calcularla es la siguiente:

$$CLV = \sum_{t=0}^{T} \frac{(p_t - c_t)\, r_t}{(1 + i)^t} - AC$$

Donde p_t es el precio pagado por el consumidor en el tiempo t, c_t es el costo directo de servir el cliente en el tiempo t, i es la tasa de descuento, r_t es la probabilidad del cliente de repetir la compra (tasa de retención, o 1 - tasa de churn), AC es el costo de adquisición del cliente, y T es el tiempo horizonte para estimar el CLV. (Kotler, P., Keller, K. L., Brady, M., Goodman, M., & Hansen, T., 2016). Si usamos un tiempo horizonte infinito, y el margen de contribución del cliente (precio menos costo) se mantiene igual, tenemos que:

$$CLV = \sum_{t=0}^{\infty} \frac{mr^t}{(1+i)^t} - AC = m\frac{r}{(1+i-r)}$$

Se ve aquí que si el flujo de fondos del cliente es mayor al costo de adquisición, valdrá la pena conseguirlo. Pero mucho más importante es ver la importancia del término r en esta ecuación. En la tabla 1 se muestra el impacto de los distintos términos de la ecuación en el Customer Lifetime Value.

Tabla 1: Impacto de los distintos términos de la ecuación en el *Customer Lifetime Value*.

Mejora de 1% en:	**Incremento en CLV**
Tasa de retención (r)	4.9%
Margen de contribución (m)	1.1%
Tasa de descuento (i)	0.9%
Costo de adquisición (AC)	0.1%

En la tabla 2 se puede ver un ejemplo de la diferencia que se puede lograr en el *Customer Lifetime Value* ante cambios en la tasa de retención y la tasa de descuento, manteniendo los márgenes fijos:

Tabla 2: Impacto del *Customer Lifetime Value* ante cambios en la tasa de retención y la tasa de descuento, para un margen fijo.

Tasa de retención (r)	**Tasa de descuento**(i)			
	10%	**12%**	**14%**	**16%**
60%	1.20	1.15	1.11	1.07
70%	1.75	1.67	1.59	1.52
80%	2.67	2.50	2.35	2.22
90%	4.50	4.09	3.75	3.46

Se ve claramente en las tablas 1 y 2 que la tasa de retención es un factor muy importante en el CLV. Una mejora de 10% en la tasa de retención puede aumentar el CLV un 33% (de 2.67 a 4.50) si se sube de 80% a 90% con un margen de contribución y una tasa de descuento fijos, mientras que las demás variables aumentan en igual o menor proporción al flujo de fondos.

En la Tabla 2 se muestra con naranja el peor escenario de los exhibidos, en donde a un mismo margen, con una tasa de retención del 60% y un tasa de descuento del 16%, un cliente tiene un *Customer Lifetime Value* igual a 1.07 del valor de su margen de contribución a la firma. Por otro lado, en verde se remarca el mejor escenario de los ilustrados en la tabla. En este contexto, con una tasa de retención de 90% y una tasa de descuento de 10%, se puede lograr un CLV de 4.5 veces el margen de contribución que aporta el cliente a la compañía. Hay que notar que ante cambios minúsculos en estos dos factores, y sobretodo en r_i, se pueden lograr cambios considerables que influyen de manera directa a las finanzas de la empresa.

Vendría muy bien para la compañía saber cuáles son las razones por las cuales los clientes se dan de baja, así poder cambiar de estrategia y poder evitar más deserciones. Mejor sería todavía saber exactamente cuáles son los clientes que se darán de baja en el próximo tiempo, para aplicar acciones específicas sobre los mismos antes que lo hagan. Con esta herramienta, la

aseguradora podría ahorrar una cantidad enorme de dinero en lugar de perder clientes sin encontrar el porqué.

1.3 - Distintos tipos de churn en seguros

Existen, a gran escala, dos maneras por las cuales un cliente puede dejar de aportar su dinero a las aseguradoras, de manera voluntaria o de manera involuntaria. Se considera fundamental explicar ambas para entendimiento del posterior análisis.

- Voluntaria: El cliente decide por sus propios medios dejar de seguir contando con los servicios de la empresa, ya sea por los precios, por baja calidad de los servicios, porque se pasó a la competencia o simplemente porque piensa que es improbable tener un accidente. Es decisión del cliente el darse de baja, y no de la compañía.
- Involuntaria: El cliente deserta por métodos de fuerza mayor. Por ejemplo muerte, mudanza, y muy comúnmente por morosidad. La decisión no es del cliente. En algunos casos, como por ejemplo el fallecimiento del cliente, la aseguradora no puede hacer nada para evitar la baja. Pero en otros casos, sí existen opciones para evitar la baja, como promociones y esquemas de pagos

1.4 - Cómo se suele abarcar

Para tratar de recortar la tasa de churn, las compañías normalmente realizan distintos métodos de seguimiento sobre los clientes mediante estadística descriptiva, como ser tableros de control y KPIs. Estos permiten generar las alertas de los clientes propensos al churn, antes que el cliente se encuentre en la fase previa del abandono. Mediante comunicaciones con el cliente, un buen marco teórico y un gran conocimiento del negocio, las firmas pueden intuir cuáles son las razones por las cuales el cliente se da de baja. También, las empresas que tratan de disminuir el ratio de deserción implementan medidas sobre los servicios para mejorar la calidad percibida,

principalmente por los clientes que renuncian a los servicios de manera voluntaria (Günther, C. C., Tvete, I. F., Aas, K., Sandnes, G. I., & Borgan, Ø., 2014).

1.5 - Propuesta de Trabajo

El propósito de este trabajo es encontrar un método basado en análisis de datos que pueda predecir de la mejor manera quiénes son los clientes que están próximos a realizar churn. La idea es generar, con ayuda de aprendizaje automático, un modelo supervisado que trate de predecir las bajas dada una cartera de clientes para un período de tiempo próximo, discriminando bajas voluntarias de involuntarias. Solo estará centrado en un tipo particular de seguro, el seguro de vida.

El modelo tratará de exponer una probabilidad de baja para el seguro de vida de cada uno de los usuarios. Para lograrlo, se probarán distintos algoritmos de machine learning para poder tener una mayor precisión. Se tratará de investigar si es mejor modelar dividiendo bajas voluntarias e involuntarias, o mantenerlas en conjunto.

Se analizará para cuánto tiempo futuro se tienen que predecir las bajas. ¿Es mejor predecir para el mes siguiente, para dentro de dos meses, o tres? La ventana de tiempo otorga la posibilidad de tomar acción antes de la baja, por ejemplo haciendo campañas de marketing mediante llamados al cliente para evitar el churn. Para esto hay que tener en cuenta el tiempo que se usa para realizar campañas sobre los clientes, y la diferencia en performance de los modelos para distintas ventanas de tiempo.

2 - Materiales y Métodos

2.1 - Datos

Los datos provienen de un banco de gran magnitud, conectado con una aseguradora muy importante, que recopila los datos mensualmente y arma distintas fuentes. Después de compilarlos por sus propios medios, todos los meses se reciben entre 12 y 13 GB de datos de la aseguradora, que están divididos en más de 30 fuentes, entre las cuales se encuentran:

- **Campañas del mes**

  Cada mes se realizan campañas de telemarketing, en donde se llaman a algunos clientes previamente seleccionados, y se los trata de convencer de que mantengan su suscripción al seguro. Algunas veces con algún tipo de promoción o descuento. De esta manera, y de acuerdo a qué tan efectivas son las campañas, se logra retener al cliente.

  Los datos sobre lo anterior son dos fuentes que contienen información de las campañas del último mes. Contienen una fila por campaña por cliente, en donde las columnas son el ID del cliente, en qué campaña estuvo, fecha de inicio y fin de la campaña, y el método por el cual se conectó con el cliente (llamado telefónico, SMS o mail). También indica para los llamados si contestó o en qué estado se encontraba, y por último para los mails indica si abrió el mail o no.[1]

[1] Los números que corresponden a los clientes fueron modificados para preservar la confidencialidad de los datos.

Tabla 3: Muestra de la fuente de campañas del mes para Enero de 2019.

ID	Envío	Open	Envio 2	Open 2	SMS	Fecha
1	SI	SI	SI	NO	NO	2019-01-19
2	SI	NO	NO	NO	SI	2019-01-19
3	SI	NO	NO	NO	NO	2019-01-19
4	SI	NO	NO	NO	NO	2019-01-19
5	SI	NO	NO	NO	NO	2019-01-19
6	SI	NO	NO	NO	NO	2019-01-19
7	SI	NO	NO	NO	SI	2019-01-19
8	SI	SI	SI	SI	NO	2019-01-19
9	SI	NO	NO	NO	NO	2019-01-19
10	SI	NO	NO	NO	NO	2019-01-19
11	SI	NO	NO	NO	SI	2019-01-19
12	SI	SI	SI	NO	SI	2019-01-19
13	SI	NO	NO	NO	NO	2019-01-19
14	SI	NO	NO	NO	NO	2019-01-19
15	SI	NO	NO	NO	NO	2019-01-19
16	SI	NO	NO	NO	NO	2019-01-19
17	SI	SI	SI	SI	NO	2019-01-19
18	SI	SI	SI	SI	NO	2019-01-19
19	SI	SI	SI	NO	SI	2019-01-19

- **Seguros**

Estas fuentes contienen los datos de los seguros del mes. Figuran por fila todas las pólizas activas hasta ese momento junto a las pólizas que se hayan dado de baja ese mismo mes. Hay 116 columnas, entre las cuales se pueden detallar, por ejemplo:

1. Tipo de orden (Alta o Baja)
2. ID cliente
3. Monto Cuota
4. Sucursal
5. Póliza anterior
6. Fecha desde préstamo
7. Estado de cuenta
8. Motivo Baja (si es que es baja)
9. Open Market o Credit Related

A continuación se muestran las primeras filas y columnas de la fuente seguros.[2]

[2] Los números que corresponden a los clientes fueron modificados para preservar la confidencialidad de los datos.

Tabla 4: Muestra de la fuente de Seguros para el mes de Enero de 2019.

ORDEN	CLIENTE	ACH_MEDIA	AMT_DEP_CURR	FEE_CURR
1953	7101232	NaN	47.55	47.55
1954	5299856	NaN	1.70	1.70
1955	4779250	NaN	1.20	1.20
1956	5665231	NaN	29.73	29.73
1957	5859611	NaN	20.14	20.14
1958	4721589	NaN	37.56	37.56
1959	6332356	NaN	34.65	34.65
1960	7100258	NaN	1.11	1.11
1961	6558960	NaN	67.40	67.40
1962	2331011	NaN	3.84	3.84

El procesamiento que se les realiza a estas fuentes es, entre otras cosas:

- Marcar las pólizas activas
- Marcar las pólizas open market y no credit related
- Se calcula la antigüedad de las pólizas, y el tiempo hasta su vencimiento
- Se generan otras variables de agregación, como ser el máximo, el mínimo, el promedio o la suma de ciertos atributos. (Staudt, M., Kietz, J. U., & Reimer, U., 1998)

Luego se organizaron en dos fuentes distintas: La primera es una fuente que tendrá solamente las órdenes de alta de seguros, y la otra tendrá las órdenes de baja de seguros. A cada una de estas también se le realiza un procesamiento. Para la fuente de altas, el procedimiento es similar a las fuentes primarias.

Para la fuente de bajas, se marcan las pólizas dadas de baja. También, si el cliente también aparece en la fuente de altas (es decir que tiene al menos un alta ese mes), la

orden se elimina. También se marca el tipo de baja según la variable de motivo de baja (que puede ser voluntario o involuntario).

- **Inversiones**

La fuente de inversiones contiene todas las inversiones que se realizaron por parte de los clientes durante cada mes. Contiene 109 columnas. Para cada inversión, se detallan, entre otros:

- El importe
- El producto
- La cantidad de acciones residuales
- Las cotizaciones
- El tipo de la especie
- La moneda de la especie

Como se ve en la tabla 5, en donde se muestran las primeras filas y las primeras columnas de la fuente, muchos valores son nulos o están mal ingresados. Habrá que realizar una depuración de los mismos.[3]

[3] Los números que corresponden a los clientes fueron modificados para preservar la confidencialidad de los datos.

Tabla 5: Muestra de la fuente de inversiones para el mes de enero de 2019.

CUS_NO	FECHA	AGENTE	CARRIER	DEP	ORIG_AMT
5056963	01-Jan-01	NaN	655	NaN	13450
5056963	01-Jan-01	NaN	725	NaN	204180
5056963	01-Jan-01	NaN	740	NaN	54458.87
5056963	01-Jan-01	NaN	859	NaN	78400
5056963	01-Jan-01	NaN	10035	NaN	273000
5056963	01-Jan-01	NaN	10042	NaN	217250
5056963	01-Jan-01	NaN	15250	NaN	55170.53
5056963	01-Jan-01	NaN	15255	NaN	247395.30
5056963	01-Jan-01	NaN	15258	NaN	1882609.15
5056963	01-Jan-01	NaN	15261	NaN	185403.30
5056963	01-Jan-01	NaN	15275	NaN	111105
5056963	01-Jan-01	NaN	15282	NaN	727912.74
5056963	01-Jan-01	NaN	30005	NaN	66500
5056963	01-Jan-01	NaN	41902	NaN	301950

- **Clientes**

Esta fuente contiene la información de cada uno de los clientes del banco asociado con la aseguradora, una fila por cliente, para el último mes. Incluye 212 columnas, entre las cuales se encuentran:

  - ID cliente
  - Edad
  - Demografía básica
  - Antigüedad
  - Variables relacionadas con inversiones
  - Variables relacionadas con créditos
  - Variables relacionadas con saldos
  - Para cada producto del banco, una variable dummy que indica si el cliente es tenedor (aquí también se ve si el cliente posee algún seguro de la aseguradora).
  - Variables relacionadas con plazos fijos
  - Variable relacionadas con fondos de inversión
  - Variables relacionadas con extracciones
  - Variables relacionadas con mora
  - Campañas en las que participó
  - Uso de canales

Tabla 6: Muestra de la fuente de clientes para el mes de Enero de 2019.

ORDEN	FECHA	CLIENTE	TITULAR	CUS_TIPO	CUS_SUBTIPO	NUEVO
1397	01-Jan-2019	80147829	0	I	I	0
1398	01-Jan-2019	80147866	0	I	I	0
1399	01-Jan-2019	80147973	0	I	I	0
1400	01-Jan-2019	80366748	0	I	I	0
1401	01-Jan-2019	80366812	0	I	I	0
1402	01-Jan-2019	80366848	0	I	I	0
1403	01-Jan-2019	80366902	0	I	I	0
1404	01-Jan-2019	80367052	0	I	I	0
1405	01-Jan-2019	80367074	0	I	I	0
1406	01-Jan-2019	80367083	0	I	I	0
1407	01-Jan-2019	80367094	0	I	I	0
1408	01-Jan-2019	80367206	0	I	I	0
1409	01-Jan-2019	80367257	0	I	I	0
1410	01-Jan-2019	80367321	0	I	I	0
1411	01-Jan-2019	817162	0	I	I	0

- **Anulaciones**

Esta fuente muestra las cancelaciones de distintos productos que tiene el banco. Se detallan las fechas de baja, la fecha de inicio del alta del producto, la fecha de finalización de contrato, la causa por la cual se anuló antes de lo previsto, el canal por el cual se realizó la anulación, entre otros. En la siguiente tabla se muestran las primeras columnas de esta fuente para dejar una idea de cómo se ve. Nota: Los números que corresponden a los clientes fueron modificados para preservar la confidencialidad de los datos.

Tabla 7: Muestra de la fuente de anulaciones para el mes de Enero de 2019.

FE_OPE	CLIENTE	RAMO	CANAL	MOTIVO	CAUSA	FE_DESDE
01-01-2019	7364666	21	SUCURSAL	52	ANULACION POR FALTA DE PAGO	12-04-2018
01-01-2019	7443628	18	SUCURSAL	47	TARJETA NO OPERATIVA	24-04-2018
01-01-2019	4710103	19	TELEMARKETING	4	CIERRE DE CUENTA	09-04-2018
01-01-2019	80760168	1	SUCURSAL	47	TARJETA NO OPERATIVA	20-04-2018
01-01-2019	60741617	19	SUCURSAL	52	ANULACION POR FALTA DE PAGO	23-04-2018
01-01-2019	71477993	19	SUCURSAL	52	ANULACION POR FALTA DE PAGO	10-04-2018
01-01-2019	71470554	19	SUCURSAL	52	ANULACION POR FALTA DE PAGO	10-04-2018
01-01-2019	50696655	26	TELEMARKETING	24	TARJETA DADA DE BAJA	12-04-2018
01-01-2019	80510374	18	SUCURSAL	14	ANULACION POR FALTA DE PAGO	12-04-2018
01-01-2019	80768927	18	SUCURSAL	47	TARJETA NO OPERATIVA	12-04-2018

- **Veraz**

Esta fuente contiene los resultado de los tests de veraz que se realizaron a distintos clientes. Una fila por test. Vemos en la siguiente tabla una muestra de la fuente Veraz. Nota: Los números que corresponden a los clientes fueron modificados para preservar la confidencialidad de los datos. En esta fuente, que es actualizada todos los meses, se muestra el resultado de la consulta de todos los clientes, no solamente las consultas que se realizaron en el mes corriente.

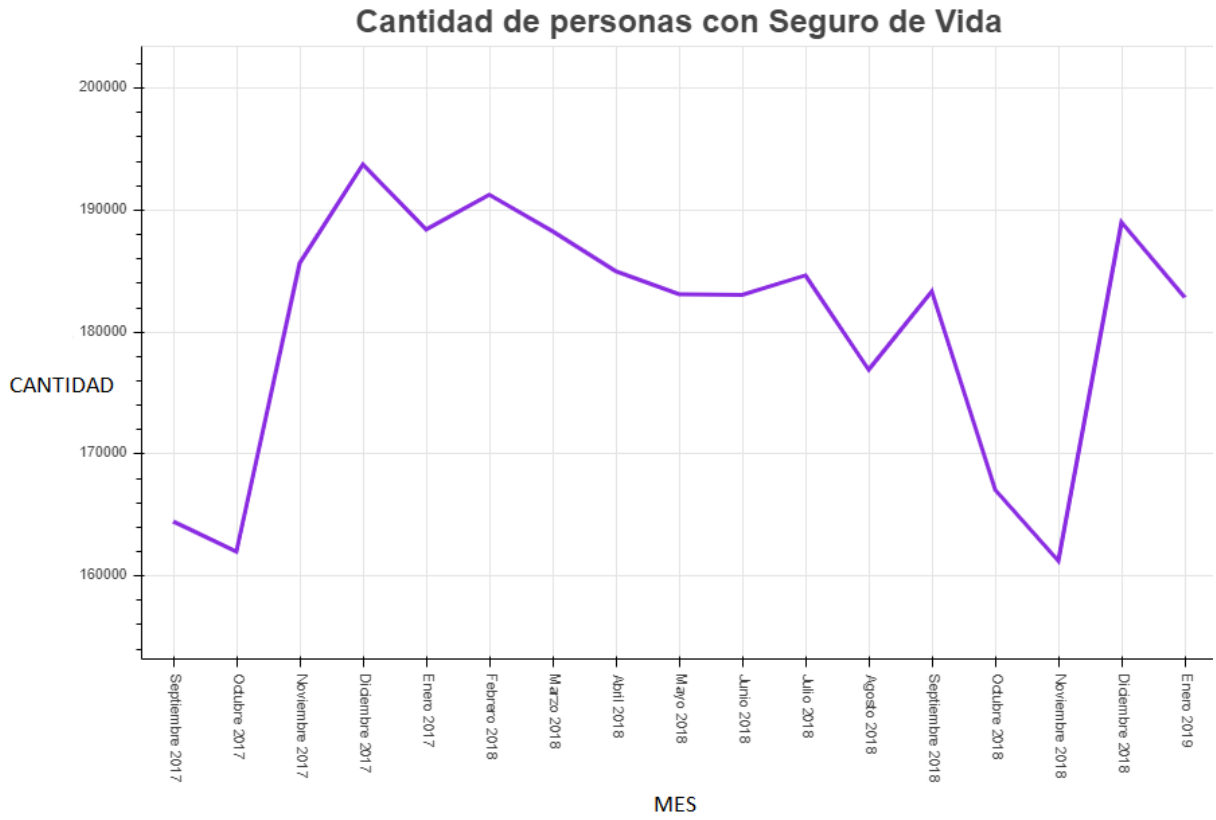
Tabla 8: Muestra de la fuente de Veraz para todas las consultas realizadas hasta el mes de Enero de 2019 inclusive.

CLIENTE	FECHA_CONSULTA	SCORE	GRUPO
64486	10-Sep-2016	NaN	NaN
64486	23-May-2015	611	C
65412	31-Jul-2012	847	Z
61280	01-Nov-2015	102	B
61280	20-Dec-2013	620	R
61280	20-Dec-2013	620	R
61280	03-Oct-2013	651	R
61280	02-Oct-2013	651	R
61280	26-Dec-2012	789	B
61280	18-May-2012	721	R

2.2 - Exploración de los datos

Actualmente se tienen datos desde enero de 2017 a enero de 2019, y todos los meses la aseguradora envía nuevos datos. Ante una primera mirada a los datos se pueden ver algunas cuestiones. En la Figura 1 se muestran las variaciones en la cantidad de clientes que poseían un seguro de vida en cada uno de los meses desde Septiembre de 2017 a Enero de 2019.

Figura 1: Cantidad de clientes de la aseguradora con seguro de vida desde Septiembre de 2017 a Enero de 2019.



En la Figura 2 se ilustran la cantidad de personas que dieron de baja su contrato voluntariamente, detallando para los meses desde Septiembre de 2017 a Enero de 2019. En la

Figura 3, para los mismos meses, se muestran por mes la cantidad de personas que dieron de baja su contrato de manera involuntaria.

Figura 2: Cantidad de anulaciones de contratos de seguros de vida voluntarios, desde Septiembre de 2017 a Enero de 2019.

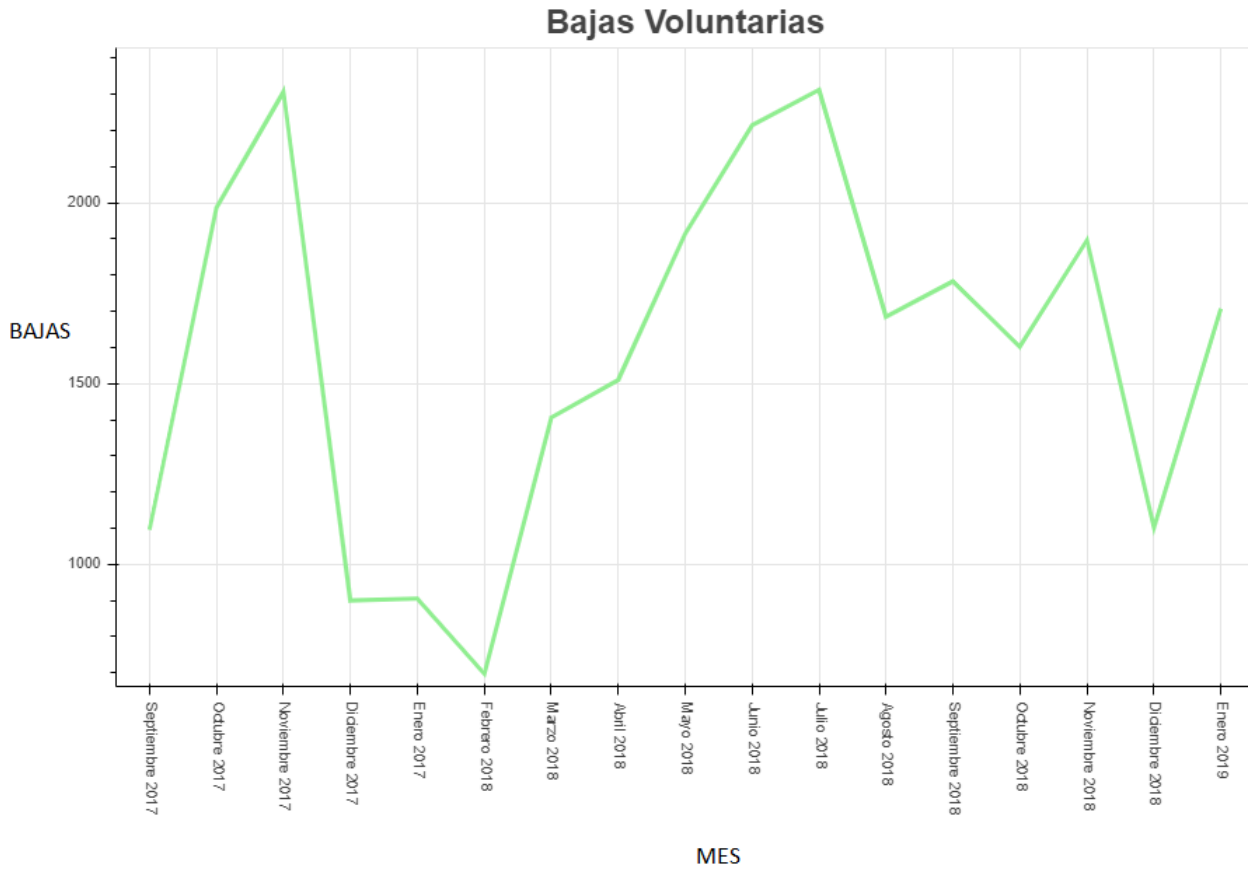
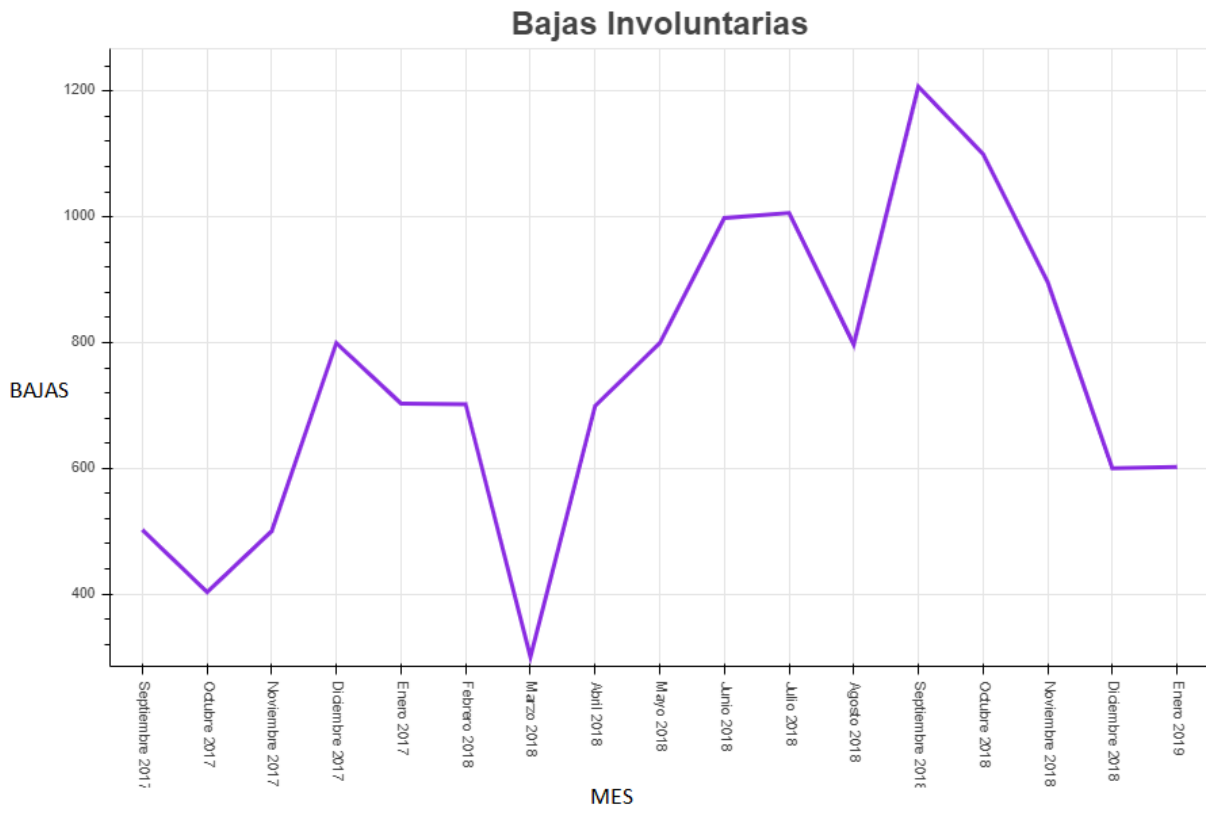


Figura 3: Cantidad de anulaciones de contratos de seguros de vida involuntarios, desde Septiembre de 2017 a Enero de 2019.

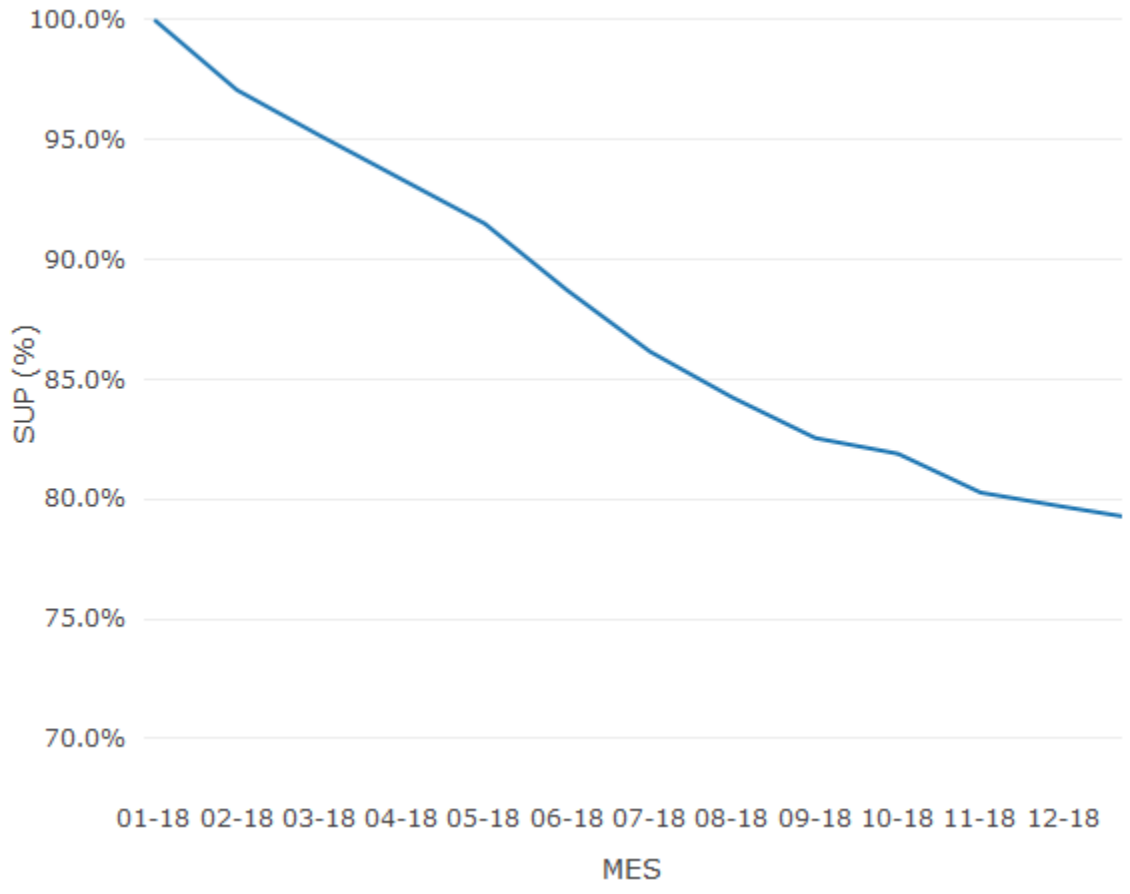


En el mes de enero de 2019, un total de 182.839 clientes tenían por lo menos un seguro de vivienda. 1707 se dieron de baja voluntariamente y 602 involuntariamente.

De los gráficos obtenidos podemos notar que no hay una evidencia fuerte que muestre una estacionalidad clara. Posibles interpretaciones de esta situación pueden ser la inestabilidad económica del país en la actualidad, y la posible correlación de la cantidad de bajas de acuerdo con la cantidad y la efectividad de las campañas para retención realizadas en diversos meses. Esto hace mucho más complicado predecir quién será el próximo en darse de baja con las reglas comerciales que utilizan los bancos y aseguradoras hoy en día.

Además, se puede observar que los clientes no mantienen su póliza mucho tiempo. En la Figura 4 se muestra, de los clientes que dieron de alta un seguro en enero de 2018, cómo varió la proporción que mantuvo el servicio para los meses subsiguientes. Solamente el 79% mantuvo la póliza un año entero. Esto resulta algo alarmante ya que indica que la tasa de retención no es óptima, y no hay una relación de fidelidad importante en el cliente. Por otra parte, se ve una leve tendencia a tener una derivada positiva, que indicaría a primera vista que los clientes que ya están hace tiempo pagando su póliza, tienden a desertar en menor proporción a los nuevos clientes.

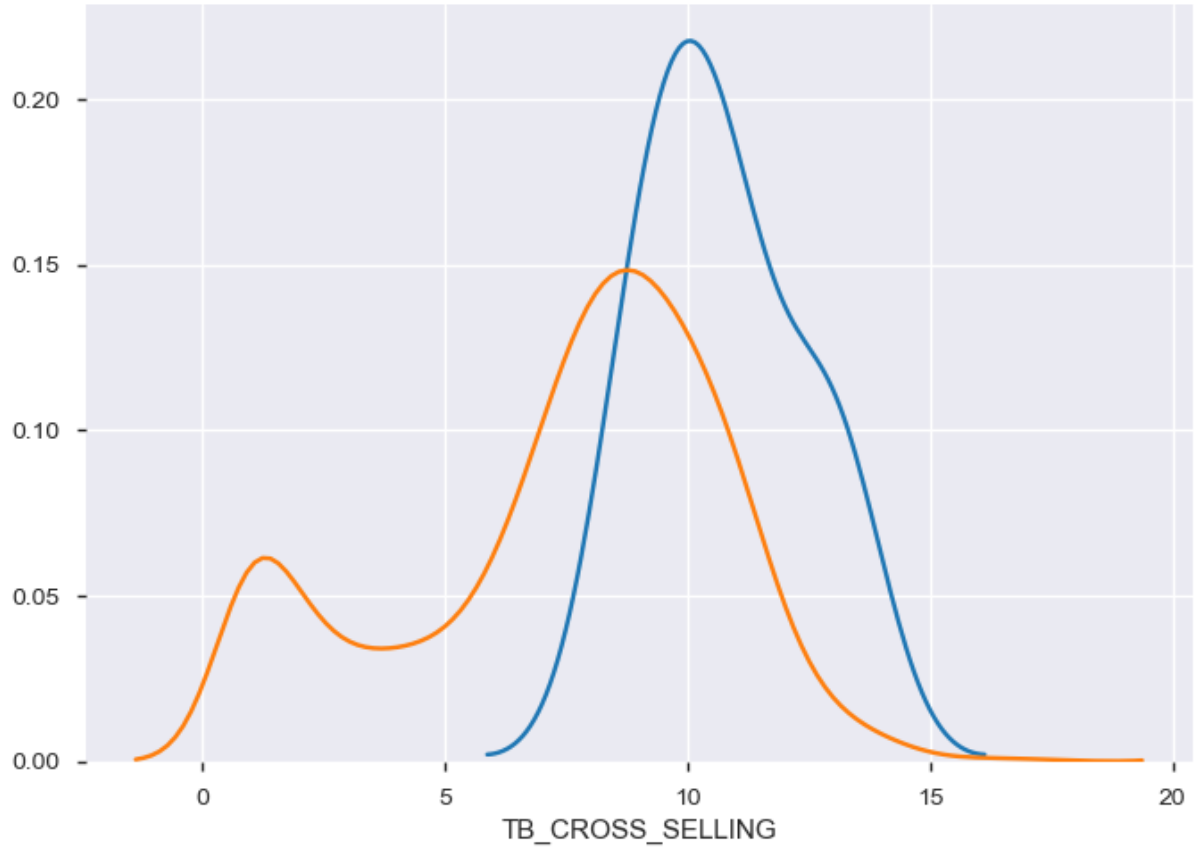
Figura 4: Curva de supervivencia de clientes que dieron de alta un seguro de vida en Enero de 2018. Muestra el porcentaje de clientes que mantienen la póliza hasta Diciembre de 2018.



Una vez depurados los datos, se pudo pasar a la siguiente etapa. El objetivo es crear un modelo no paramétrico que sea capaz de predecir las bajas mediante el hallazgo de patrones entre variables predictivas y la variable dependiente, que sería justamente la baja del usuario.

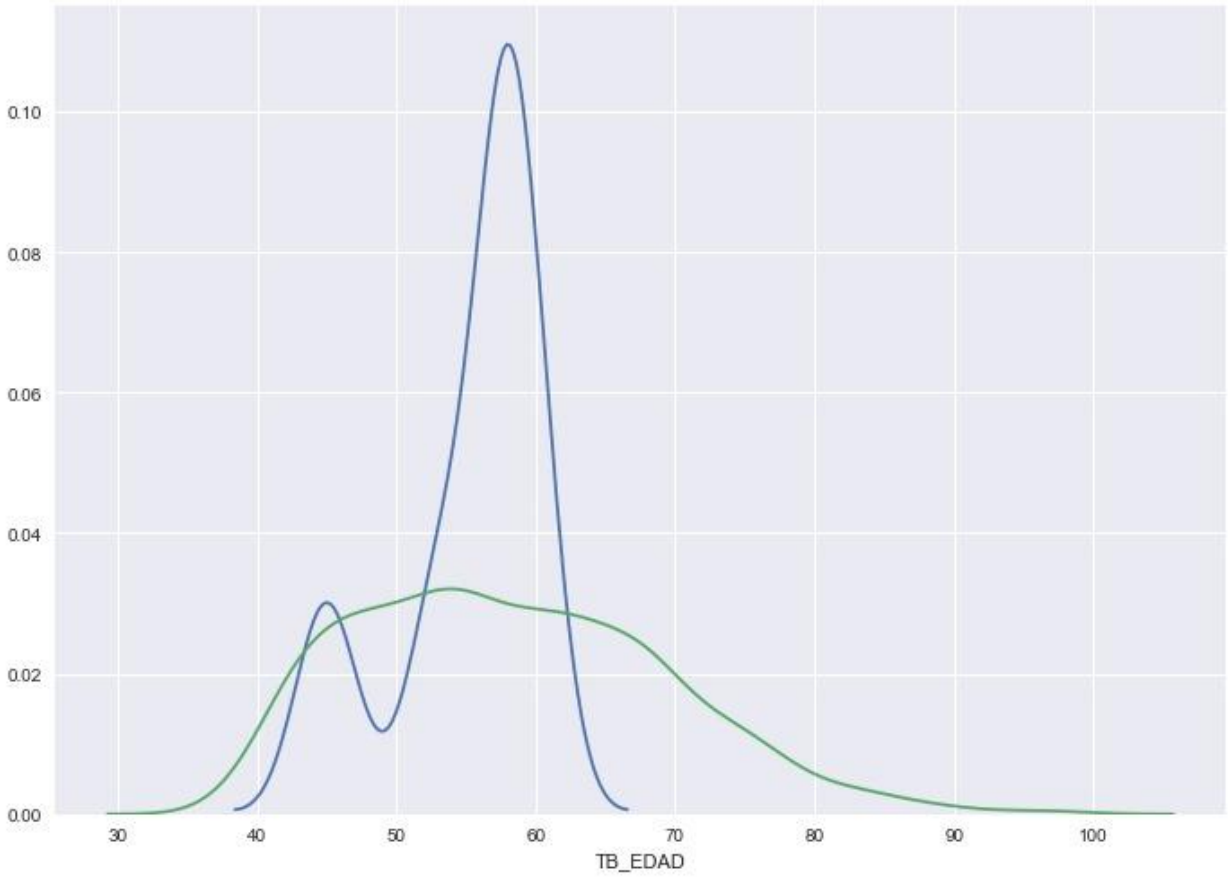
Por ejemplo, podemos ver algunas correlaciones a ojo. En la Figura 5, se muestran las estimaciones de la densidades de Kernel de la variable TB_CROSS_SELLING (variable que indica la cantidad de productos que posee el cliente en el momento) de acuerdo a si el cliente se da de baja o no. La distribución azul contiene todos los clientes que se dieron de baja en un mes en particular (en este caso estamos viendo el mes de enero de 2019), mientras que la distribución naranja muestra los clientes que mantuvieron su póliza activa durante ese mes. Se nota una diferencia entre ambas distribuciones, que podría dar a pensar que mientras más productos tiene el cliente en su poder, más probabilidades hay de que el cliente dé de baja un seguro de vida.

Figura 5: Estimación de la distribución de Kernel para la variable TB_CROSS_SELLING. La distribución naranja correponde a los usuarios que mantuvieron su póliza en enero de 2019, mientras que la distribución azul corresponde a los usuarios que dieron de baja su póliza el mes de enero de 2019.



También, si se analizan las distribuciones de los clientes que se mantuvieron el seguro y los que se dieron de baja, se puede ver que hay muy pocos clientes por arriba de los 65 años que dieron de baja la póliza, como muestra la Figura 6. Esto permite inducir que a medida que la gente se hace mayor, deja de importarle tanto la cuota del seguro porque valora más su vida, o simplemente porque quiere estar asegurado ante cualquier eventualidad, a la que intuye una probabilidad cada vez mayor con el paso del tiempo. (Guillén, M., Nielsen, J. P., Scheike, T. H., & Pérez-Marín, A. M., 2012).

Figura 6: Estimación de la distribución de Kernel para la variable TB_EDAD. La distribución verde correponde a los usuarios que mantuvieron su póliza en enero de 2019, mientras que la distribución azul corresponde a los usuarios que dieron de baja su póliza el mes de enero de 2019.



Si con las dos variables anteriores se puede generar un entendimiento bastante amplio de cómo se diferencian los clientes que dan de baja un seguro de vida contra los que no lo hacen, ¿cómo afectaría poder hacer un análisis con más de 2000 variables? Esto es algo que el ser humano mediante reglas de marketing no puede realizar, al menos en un tiempo óptimo para poder inferir conclusiones rápidas. Es por eso que en este trabajo se plantea el aprendizaje automático para que la computadora aprenda por sí misma y pueda llegar a diferenciar de una buena manera a estos dos tipos de clientes.

2.3 - Modelos a Usar

Antes de comenzar a experimentar, es necesario planificar la estructura de los modelos que se compararán, definir la métrica con la cual se contrastarán los desempeños, y definir otros detalles. Los parámetros para los cuales se evaluarán los modelos no solo dependen de cuál modelo se utilice, sino también de cómo repartiremos los datos entre entrenamiento y validación, cuál es la ventana de tiempo a evaluar, la cantidad de meses anteriores a considerar, y el balanceo de los datos.

Primero tendríamos que evaluar si el modelo tendría que ser distinto para voluntarios que para involuntarios. Es decir, un modelo de clasificación múltiple en donde el output sea si no se dio de baja, se dio de baja voluntariamente, o se dio de baja involuntariamente. La otra opción es crear dos modelos por separado, en donde cada uno prediga si el cliente se da de baja o no, (si el cliente se da de baja voluntariamente, aparecerá como que se da de baja en el modelo de voluntarios pero no en el de involuntarios). (Berry, M. J., & Linoff, G. S., 2004). En la Figura 7 se muestra un ejemplo de cómo quedaría la variable dependiente de acuerdo a si se realiza un modelo de clasificación múltiple o dos modelos de clasificación binario.

Figura 7: Muestra de cómo quedaría la variable dependiente de acuerdo a si se realiza un modelo de clasificación múltiple o dos modelos de clasificación binario.

MODELO CLASIFICACIÓN MULTICLASE

VAR 1	VAR 2	VAR 3	...	VAR N	BAJA
					NO
					NO
					VOLUNTARIA
					NO
					NO
					INVOLUNTARIA

DOS MODELOS CLASIFICACIÓN BINARIA

VAR 1	VAR 2	VAR 3	...	VAR N	BAJA VOLUNTARIA
					NO
					NO
					SI
					NO

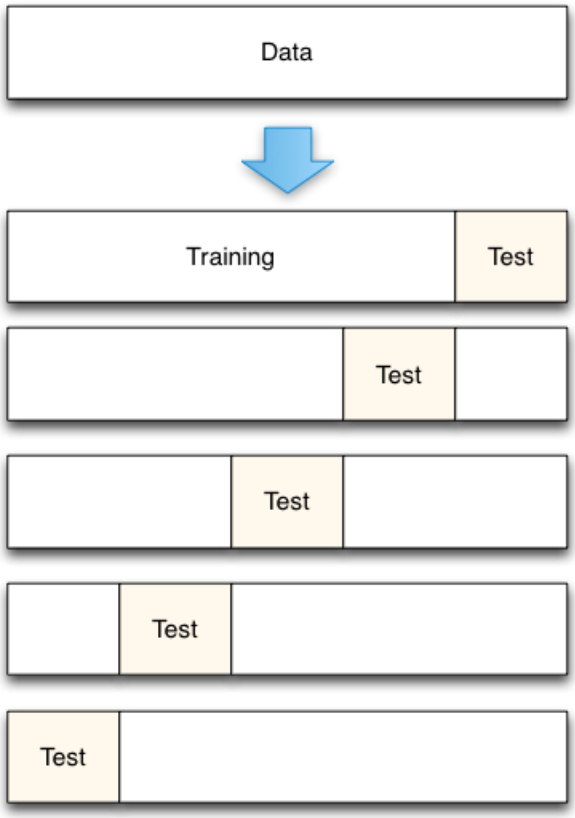
VAR 1	VAR 2	VAR 3	...	VAR N	BAJA INVOLUNTARIA
					NO
					NO
					NO
					NO
					NO
					SI

Para lograr esto, se deberían presentar de distinta manera los datos. Si se realiza un modelo de clasificación múltiple, solamente dejando un dataset con todas las variables y una fila

por cliente, estaría en buenas condiciones de ser entrenado. En caso de realizar por separado dos modelos de clasificación binaria; un modelo para voluntarios y otro para involuntarios, en el modelo de voluntarios, si un cliente se dio de baja involuntariamente, aparecerá como que no desertó. Lo mismo pasará en el modelo de involuntarios. Se detalla esto en la Figura 7.

Para poder elegir cuál es el mejor modelo, habría que medirlos de alguna manera. Se proponen dos formas distintas. La primera es con validación cruzada de 5 iteraciones. Cross Validation es una buena manera de calcular la *performance* de las predicciones, ya que en cada iteración se elige un set de datos aleatorio distinto con muestreo aleatorio simple mientras se entrena el modelo con los restantes datos, y se calcula la performance promedio. De esta manera podemos eliminar los posibles sesgos de selección al contrastar dos modelos en donde uno puede estar sobre-ajustándose a los datos.(James, G., Witten, D., Hastie, T., & Tibshirani, R., 2013).

Figura 8: Muestra de cómo se dividen los datos en un método de validación cruzada (*Cross Validation*). Cada partición tiene observaciones elegidas al azar con muestreo aleatorio simple-



La otra manera de calcular la calidad del desempeño de los modelos es mediante su implementación en tiempo real. Se aplicarán predicciones para 3 o 4 meses después del mes donde se tienen los últimos datos, y se compararán las métricas contra ese período. Esto nos importa más ya que es la verdadera implementación práctica que se le dará a los modelos, y en donde se podrá ver realmente qué tan efectivos son.

Los datos claramente están desbalanceados. Por lo tanto, puede resultar malo para los modelos clasificadores ya que, al tratar de minimizar error, es más fácil predecir que ningún cliente se dará de baja del seguro. Un modelo trivial que prediga que ningún cliente se dará de baja, tendrá una buena *performance* en *accuracy*, pero no nos servirá para los fines de este trabajo.

Debido a la gran cantidad de datos y las limitaciones de la memoria RAM, para poder correr los modelos en la computadora fue necesario generar una muestra aleatoria de los mismos. Esta técnica puede perder información valiosa ya que se descartan muchas filas. (Visa, S., & Ralescu, A., 2005). Para tratar de perder la menor información posible, se eliminó cierta cantidad al azar de muestras que no se hayan dado de baja en el dataset a entrenar, manteniendo las filas en donde el cliente efectivamente realizó churn. Al realizar esto, después se tuvieron que tener en cuenta algunas consideraciones, por ejemplo al momento de validar la performance de los mismos. Una técnica que se utiliza en modelos basados en árboles es el ponderar el hiperparámetro que indica el peso de las clases positivas (scale_pos_weight). Esta técnica puede ayudar a mitigar el problema. Cuando se realizó *K-Fold Cross Validation*, se tuvo en cuenta el ratio de las muestras que se borraron para poder ingresarlos ponderados en el parámetro de la función set_weight de la librería LightGBM para Python. Para hacer una competencia más justa entre los modelos, se decidió que todos los modelos compitan con el mismo set de datos. (Japkowicz, N., 2003).

En cuanto a los valores faltantes que no se pudieron imputar, se los modificó por un valor fijo de -30. Normalmente LightGBM maneja bien los *missings*, pero se eligió optar por esta alternativa al ver que no había cambios significativos entre un procedimiento y otro.

Hay que tener cuidado sobre la métrica con la que se mide la performance de los modelos. Si se utiliza la accuracy (exactitud), en donde la calidad del modelo depende del porcentaje de aciertos, puede llevar a que los modelos predigan que no existe el churn al haber tantos datos desbalanceados. Con tasas tan bajas, predecir que no va a haber gente que deje de consumir el seguro daría una accuracy muy alta (exactamente del 1 - tasa de churn). Se debería buscar una métrica en donde se premie al encontrar alguien que efectivamente se da de baja. (Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H., 2006).

El área bajo la curva ROC (AUROC) es una buena métrica para este tipo de modelos de clasificación binaria, en donde se mide la probabilidad de que el modelo puntúe una instancia positiva aleatoria más alta que una negativa.(Fawcett, T., 2006). La curva ROC se define graficando la tasa de verdaderos positivos en el eje y, y la tasa de falsos positivos en el eje y. Para este trabajo

en particular, se definió como métrica el coeficiente GINI, que es simplemente una transformación lineal de el área bajo la curva ROC (GINI = 2 x AUROC - 1). El coeficiente GINI tiene un rango desde 0 a 1, siendo 0 un modelo totalmente aleatorio, y 1 un modelo que discrimina perfectamente quién se da de baja y quién no lo hace. Mientras más alto el GINI, mejor predice el modelo.[4]

Ingeniería de Variables

Una de las medidas que se tomaron fue la de realizar ingeniería de variables. En concreto lo que se realizó fue la agregación de todas las variables para todos los períodos en los que se tienen datos. Se calcularon, para cada cliente, la suma, la media, el máximo, el mínimo y el desvío estándar para los últimos 3, 6 y 9 meses dependiendo la cantidad de períodos. Es importante realizar esto ya que ayuda a las limitaciones que tienen los modelos basados en árboles para realizar agregaciones por sí mismos (Chen, T., & He, T., 2015). Por ejemplo, si un árbol necesita calcular si un cliente estuvo en mora durante los últimos 3 meses, debería realizar por lo menos 3 cortes, en lugar de uno solo si ya se le brinda esta variable desde el inicio.

LightGBM

LightGBM es un modelo relativamente nuevo, que aplica el gradient boosting a algoritmos basados en árboles. Se eligió este algoritmo como alternativa ya que el dataset final tiene muchas filas y este algoritmo aplica muy bien para grandes escalas de datos. (Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y., 2017). A pesar de que es más difícil de implementar que otros tipos de algoritmos de árboles como decision trees y random forest, generalmente obtiene mejores resultados. Se utilizó la librería LightGBM, en particular la versión 2.2.2, para Python 3.7.1 para poder realizar los modelos.

[4] Para el cálculo de GINI para modelos de clasificación multiclase, se utilizó la función exponencial normalizada, o *softmax*, que se encuentra en la librería LightGBM si se fija como parámetro *objective : 'multiclass'*.

Para la optimización de hiperparámetros, se hizo uso de la técnica de *Random-Search* con 5 distintos hiperparámetros, entre ellos:

- Learning Rate : [0.01, 0.04]
- Leaves: [48, 80]
- Min Child Weight: [10, 40]
- Col Sample by Tree: [0.25, 0.75]
- Number of iterations: [300, 700]

Random-Search es una técnica que se utiliza para la optimización de hiperparámetros de los distintos modelos, en donde en lugar de probar todas las combinaciones posibles, como se haría en *Grid-Search*, se prueban combinaciones aleatorias de los distintos hiperparámetros para poder llegar al mejor modelo posible. Es decir, se eligen aleatoriamente un set de coeficientes para cada uno de los hiperparámetros, y luego se calcula la performance. Esto se realiza una cantidad limitada de iteraciones, y generalmente se pueden llegar a encontrar resultados no alcanzables con *Grid-Search* con una cantidad limitada de iteraciones.

Para los siguientes resultados se tuvieron en cuenta 30 iteraciones de random-search con lo parámetros denotados arriba para cada una de las combinaciones siguientes:

- Cantidad de períodos a entrenar (3, 6 o 9 meses)
- Ventana de tiempo (1 o 2 meses)
- El modelo que se realizó (juntando voluntarios e involuntarios en un solo modelo, o haciendo dos modelos distintos)

Es decir, para cada una de las filas en la tabla, se realizó un random-search de 30 iteraciones. Por lo tanto, si queremos hacer la cuenta de cuántos modelos se realizaron para probar el mejor, tenemos 3x2x2x30 = 360 modelos.

2.4 - Ventana de tiempo

Actualmente, la consultora recibe los datos de la compañía un mes y medio después. Es decir, en marzo de 2019 la consultora recibe los datos de enero de 2019. Esto nos da un margen menor para predecir, ya que seguramente al momento de recibir los datos, ya hay clientes que se dieron de baja y no se pudo actuar sobre ellos. El desafío es tratar de predecir quiénes son los más propensos a darse de baja en el plazo lo más corto posible.

Tabla 9: Alternativa de calendario de planificación para una ventana de 2 meses entre la predicción y las bajas efectivas

Abril	**Mayo**	**Junio**	**Julio**
A fin de mes se reciben los datos de marzo	Se entrenan los modelos con los datos de marzo. A principio de mes se entregan predicciones. Comienzo de la acción.	Mes de acción	Se dan de baja

TOTAL: 4 meses de diferencia entre los datos y el mes a predecir.

Tabla 10: Alternativa de calendario de planificación para una ventana de 1 mes entre la predicción y las bajas efectivas

Abril	**Mayo**	**Junio**
A fin de mes se reciben los datos de marzo	Se entrenan los modelos con los datos de marzo. A principio de mes se entregan predicciones. Comienzo y fin de la acción.	Se dan de baja

TOTAL: 3 meses de diferencia entre los datos y el mes a predecir.

Hay un trade-off en elegir la ventana óptima de tiempo. En principio, elegir una ventana de tiempo lejana permite tener más tiempo para accionar contra los clientes propensos a darse de baja mediante campañas de marketing hacia los mismos, descuentos, beneficios, etc.

Pero esto a su vez influye en las predicciones. Vemos a medida que aumenta la distancia entre los meses con los que se entrenó el modelo y el mes que se quiere predecir, puede existir un concept drift muy importante en donde un mes más puede disminuir mucho las performances de los modelos. (Domeniconi, C., Perng, C. S., Vilalta, R., & Ma, S., 2002). Es por esto que vemos necesaria la implementación óptima de la ventana de tiempo para la cual evaluaremos los resultados. Las dos opciones que se evaluarán serán predecir bajas para 4 meses después del último mes con datos (mostrado en la tabla de arriba), y la segunda para 3 meses (en donde se el tiempo de acción hacia los clientes propensos a la baja se reduciría en un mes para poder obtener mayor performance en las predicciones). Los resultados se muestran en la siguiente sección.

3 - Resultados

Los siguientes resultados muestran la performance medida en la métrica GINI para la mejor combinación de hiperparámetros para cada modelo. Es decir, para cada una de las filas en las tablas, se realizó un random-search de 30 iteraciones.

Tabla 11: Resultados de la optimización de hiperparámetros para los modelos probados en bajas voluntarias. Se muestran las métricas GINI en Cross Validation y GINI en el mes en el cual se predicen las bajas (Out of Time).

Períodos	Ventana	Modelo	GINI Cross Validation	GINI OOT
3	1	V	32.62%	30.88%
6	1	V	36.80%	30.30%
9	1	V	38.89%	**32.76%**
3	1	I y V	-	24.25%
6	1	I y V	-	25.30%
9	1	I y V	-	25.71%
3	2	V	32.28%	22.37%
6	2	V	36.24%	24.12%
9	2	V	36.92%	24.46%
3	2	I y V	-	19.48%
6	2	I y V	-	20.28%

Viendo los modelos que predicen para bajas involuntarias:

Tabla 12: Resultados de la optimización de hiperparámetros para los modelos probados en bajas voluntarias. Se muestran las métricas GINI en Cross Validation y GINI en el mes en el cual se predicen las bajas (Out of Time).

Períodos	**Ventana**	**Modelo**	**GINI Cross Validation**	**GINI OOT**
3	1	I	85.59%	79.60%
6	1	I	85.53%	81.82%
9	1	I	86.36%	**82.20%**
3	1	I y V	-	76.57%
6	1	I y V	-	76.97%
9	1	I y V	-	78.37%
3	2	I	80.47%	70.90%
6	2	I	80.97%	70.97%
9	2	I	80.11%	71.83%
3	2	I y V	-	66.98%
6	2	I y V	-	65.44%

Nota: las métricas de GINI Cross Validation para modelos combinados no están reportados porque la librería LightGBM entrega el resultado de la métrica para el modelo de clasificación multiclase, y resultaría injusto comparar con los otros modelos de clasificación binaria.

Por lo que se puede observar arriba, en todos los casos tener modelos separados para bajas voluntarias y bajas involuntarias es mejor para la performance en OOT (Out of Time). Esta

performance se la calcula con las bajas con las que se implementaría el modelo. Es decir, si un modelo predice las bajas de marzo con los datos de enero, una vez que se tengan los datos de marzo se podrá calcular la *performance* definitiva del modelo. A esto lo llamamos performance en *Out of Time*. Se ve por ejemplo, en la tabla de resultados de bajas involuntarias, ante un modelo de clasificación binaria con una ventana de tiempo de 2 meses y con 6 períodos de datos, una performance en GINI OOT de 70.97%. Para un modelo combinado de clasificación múltiple, pero utilizando la misma cantidad de períodos y la misma ventana de tiempo, el GINI OOT es de 65.44%. Esto ocurre en todos los casos, por lo tanto se puede afirmar que utilizar dos modelos de clasificación binaria es una buena idea para este caso.

Como era de esperar, la performance en OOT disminuye notablemente cuando la ventana de tiempo para actuar es de dos meses en lugar de una. Por ejemplo, para un entrenamiento con 3 períodos, un modelo específico para bajas voluntarias tiene una performance de 79.6% en GINI Out of Time, contra un 70.9% de GINI para el mismo modelo pero con una ventana de tiempo de dos meses. Esto refuerza la intuición de que entrenar con datos más actuales mejora el poder de predicción, y por lo tanto es fundamental tratar de manejar información más reciente.

También, entrenar con más períodos mejora la métrica GINI en OOT, aunque no con tanta fuerza como los otros parámetros. Se evidencia cuando se comparan modelos de la misma familia, con una ventana de tiempo similar. Como es el caso del 24.46% de GINI OOT en un modelo de clasificación binario para bajas voluntarias con una ventana de tiempo de 2 meses entrenado con 9 meses de historia. Se contrasta con un 24.12% y un 22.37% de GINI OOT para exactamente el mismo modelo, solamente que entrenados con 3 y 6 meses de historia respectivamente.

Por último, se ve una clara diferencia entre la performance de los modelos cuando se predicen bajas involuntarias y cuando se predicen bajas voluntarias. Esto es intuitivo, ya que las bajas involuntarias generalmente se dan debido a problemas crediticios o moras, reglamentado por la aseguradora. Se ve por ejemplo, que un modelo para voluntarios con dos meses de ventana de tiempo entrenado con 9 meses de datos anteriores, obtiene una métrica GINI de 71.83% si se predicen bajas involuntarias. Si se realiza el mismo modelo para bajas voluntarias, después de la

optimización de los hiperparámetros del LightGBM se llega a una métrica GINI de 24.46% solamente.

Se eligieron los modelos que más puntaje obtuvieron en la métrica GINI OOT. Justamente son los modelos que fueron entrenados con 9 meses de historia, discriminando modelos para voluntarios e involuntarios, y con una menor ventana de tiempo. Las métricas GINI fueron de 32.76% para el modelo de bajas voluntarias, y de 82.20% para el modelo de bajas involuntarias.

Tabla 13: Estadística predictiva de las probabilidades predichas por los modelos elegidos en enero de 2019 para predecir bajas de Marzo de 2019.

	Voluntarios	**Involuntarios**
Media	1,02	1,37
Desvío Estándar	0,86	4,96
Mínimo	0,11	0,004
Cuartil 1	0,61	0,07
Cuartil 2	0,82	0,18
Cuartil 3	1,16	0,67
Máximo	40,76	89,71

La tabla anterior muestra una estadística descriptiva de los resultados predichos para los modelos específicos (es decir, voluntarios e involuntarios por separado) un mes en particular (se anonimizaron los datos para preservar la confidencialidad de la aseguradora. Los resultados están medidos en porcentajes (La media de Involuntarios se interpreta como 1,37%). Se nota que más de la mitad de las predicciones son probabilidades menores al 0,2% para involuntarios, y 0,82% para voluntarios. Pero a su vez, se muestra que las probabilidades explicitadas para las bajas voluntarias

tienen una dispersión menor que las probabilidades explicitadas para las bajas involuntarias. Esto se muestra en el desvío estándar de cada una de las columnas (el desvío en voluntarias es menor). También podemos ver que el máximo en bajas involuntarias es contundentemente mayor al máximo de bajas voluntarias, indicando que el modelo de bajas involuntarias tiene más confianza al predecir una deserción que el de bajas voluntarias. Esto indicaría a primera instancia que los modelos de churn involuntarios tienen mayor performance que los modelos voluntarios. Abajo se muestran dos gráficos de distribución, mostrando la proporción de probabilidades de cada modelo.

Figura 9: Distribución de las probabilidades predichas por el modelo para bajas voluntarias de Marzo de 2019.

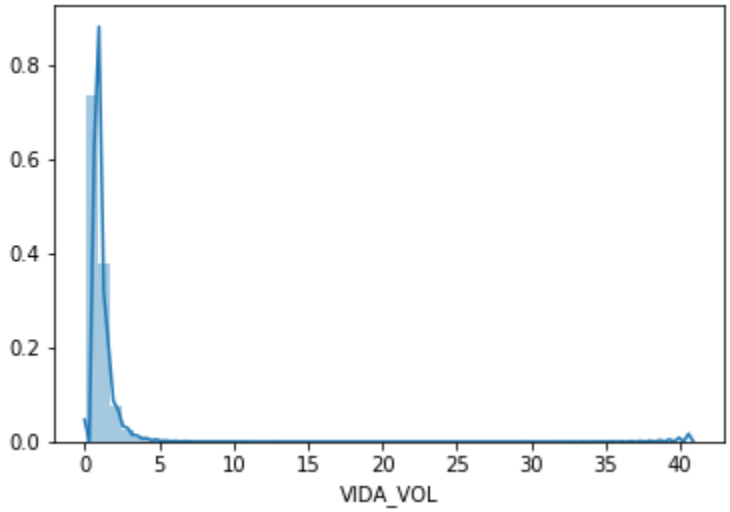
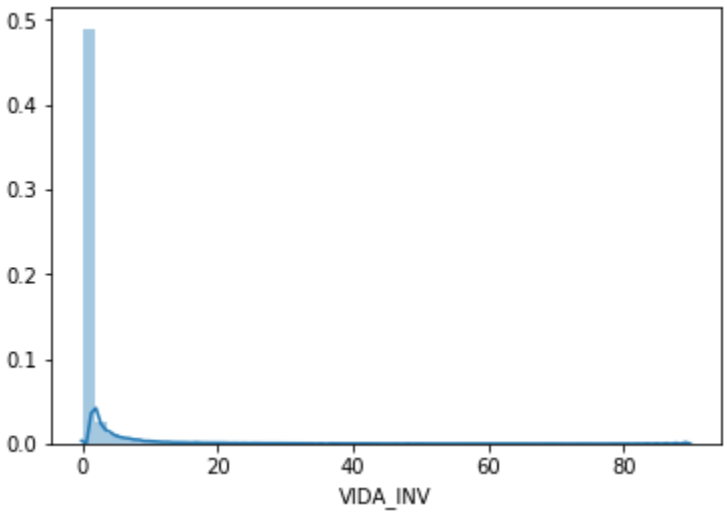


Figura 10: Distribución de las probabilidades predichas por el modelo para bajas involuntarias de Marzo de 2019.



3.1 - Desempeños

Una vez elegida la ventana de tiempo, y la cantidad de meses a entrenar, se volvió a entrenar el modelo en distintos períodos de tiempo con distintos hiperparámetros, teniendo como datos a los meses inmediatamente anteriores al mes que se quiere predecir menos la ventana de tiempo. La búsqueda de parámetros siempre se realiza con random-search para poder predecir en cada mes. Esto es importante ya que los datos van cambiando con el tiempo, y los hiperparámetros óptimos para un mes pueden no ser los óptimos para meses consiguientes.

En los gráficos siguientes se muestran los resultados de la performance del modelo medida con la métrica GINI, junto con los resultados de la performance del modelo fuera de tiempo (OOT) que se muestran con línea punteada, para poder notar cómo varía la calidad de las predicciones en cross validation y en el mes que se quiere predecir efectivamente.

Involuntarios:

Figura 11: Desempeño de los modelos para bajas involuntarias entrenados con Cross Validation y en Out of Time, desde el mes de Diciembre de 2017 hasta diciembre de 2018.

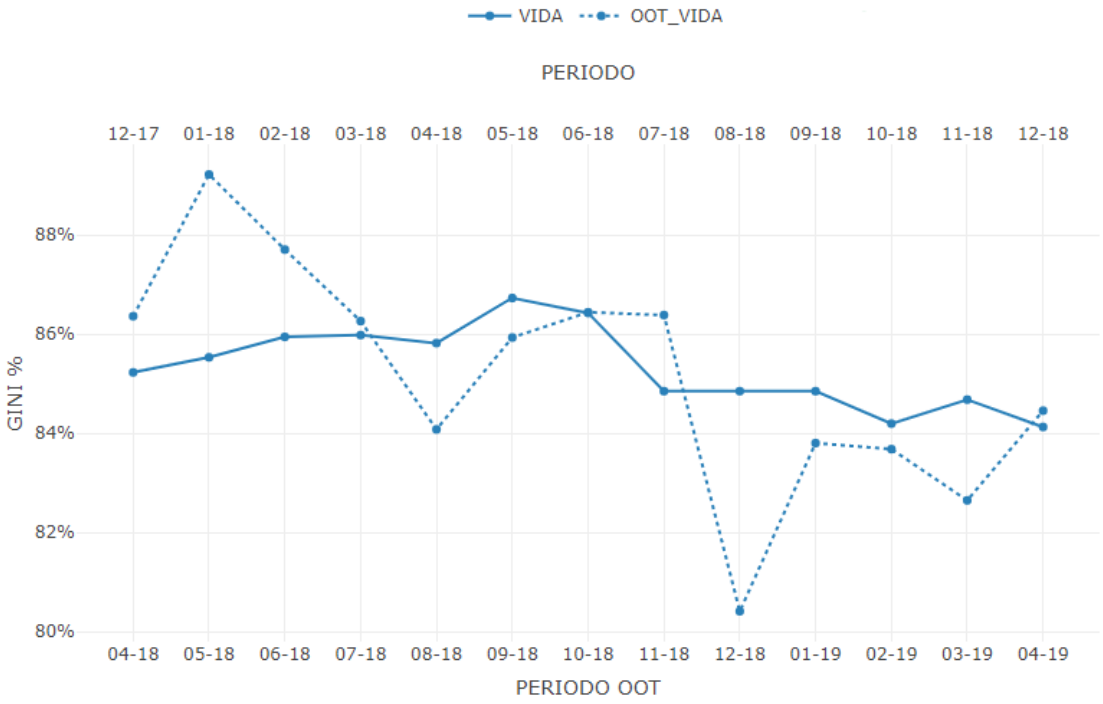


Tabla 13: Desempeño de los modelos para bajas involuntarias entrenados con Cross Validation y en Out of Time, desde el mes de Diciembre de 2017 hasta diciembre de 2018.

Fecha	GINI Cross Validation	Fecha OOT	GINI OOT
12/17	85.23%	04/18	86.36%
01/18	85.53%	05/18	89.22%
02/18	85.94%	06/18	87.70%
03/18	85.98%	07/18	86.26%
04/18	85.82%	08/18	84.08%
05/18	86.73%	09/18	85.93%
06/18	86.42%	10/18	86.44%
07/18	86.50%	11/18	86.38%
08/18	84.63%	12/18	80.41%
09/18	84.85%	01/19	83.80%
10/18	84.19%	02/19	83.68%
11/18	84.72%	03/19	82.62%
12/18	84.09%	04/19	84.46%

Voluntarios:

Figura 11: Desempeño de los modelos para bajas voluntarias entrenados con Cross Validation y en Out of Time, desde el mes de Diciembre de 2017 hasta octubre de 2018.

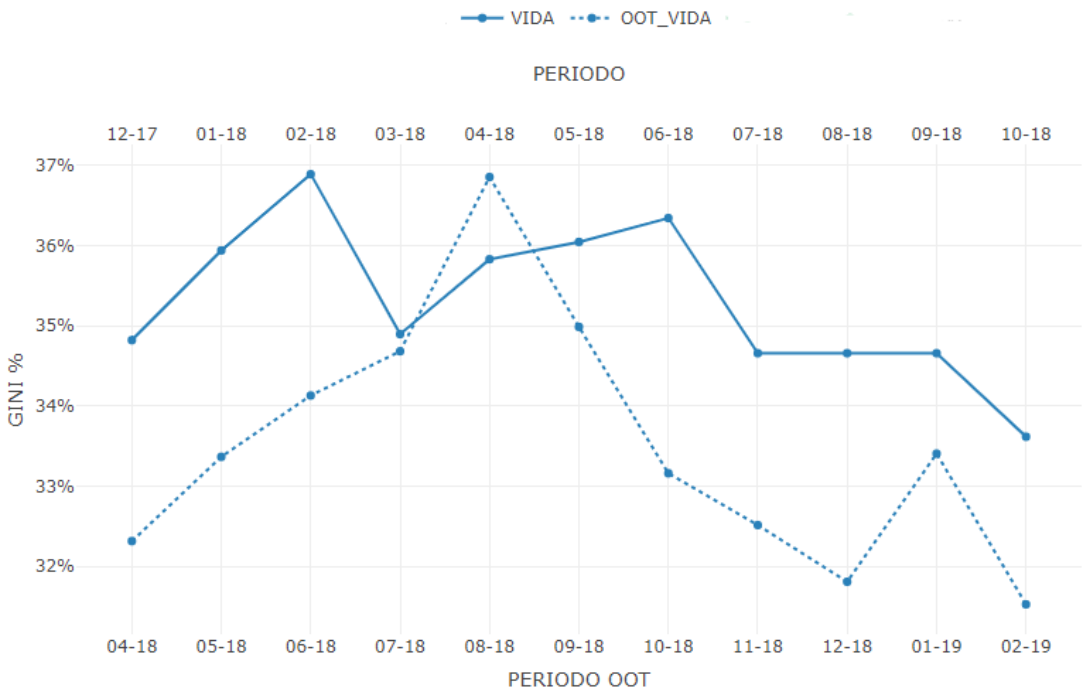


Tabla 14: Desempeño de los modelos para bajas voluntarias entrenados con Cross Validation y en Out of Time, desde el mes de Diciembre de 2017 hasta octubre de 2018.

Fecha	GINI Cross Validation	Fecha OOT	GINI OOT
12/17	34.82%	04/18	32.32%
01/18	35.94%	05/18	33.37%
02/18	36.89%	06/18	34.13%
03/18	34.90%	07/18	34.68%
04/18	35.83%	08/18	36.85%
05/18	36.05%	09/18	34.99%
06/18	36.34%	10/18	33.16%
07/18	36.01%	11/18	32.52%
08/18	34.25%	12/18	31.81%
09/18	34.66%	01/19	33.41%
10/18	33.62%	02/19	31.53%

Para el modelo de involuntarios, notamos que hay un cierto período de tiempo (desde abril de 2018 hasta julio del mismo año) en donde la performance out of time es mayor a la que hubo en cross-validation. Esto puede estar vinculado con la baja en la cantidad de altas que hubo en el mes de marzo de 2018. Vemos que la performance en cross-validation es bastante estable, y que se puede mantener en el tiempo. Hay que notar que la performance es estable porque se realiza un re-entrenamiento de los modelos, con optimización de hiperparámetros del LightGBM para todos los meses, ya que los hiperparámetros óptimos pueden cambiar de acuerdo a los datos.

Para los modelos de voluntarios, además de ver en promedio un coeficiente GINI considerablemente menor al de los modelos de bajas voluntarias, vemos una mayor dispersión en

las performance con el paso del tiempo. Esto está relacionado ya que al haber menor cantidad de bajas acertadas por el modelo, un acierto puede cambiar notablemente el rendimiento del mismo.

3.2 - Importancia de variables

Viendo los resultados de los modelos, a continuación se intentará encontrar un poco de interpretabilidad a los mismos. Saber cuáles fueron las variables que más influyeron a la hora de predecir las potenciales bajas. Es por eso que se hizo un análisis de importancia de variables, usando la métrica que utiliza LightGBM para calcularlas. Esta métrica, que se encuentra en la documentación de la librería, cuenta cuántas veces se utilizó la variable en el modelo. Mientras más usada, más importante.

En las tablas 15 y 16 se agrupó a las variables según la fuente, para poder ver cuáles fuentes son las más influyentes. Agregamos la fuente INGENIERÍA DE VARIABLES en donde se ubican todas las variables creadas debido a la ingeniería de variables. Hay que aclarar que se normalizaron los coeficientes de importancia para que la suma sea igual a 1, y poder comparar bien entre ellas.

Tabla 15: Agrupación por fuentes de la importancia de variables para el modelo de bajas voluntarias para el mes de Enero de 2019.

Fuente	Importancia Variables Voluntarios
INGENIERÍA DE VARIABLES	0.620614
SEGUROS	0.21694
CLIENTES	0.132484
VERAZ	0.011952
VERAZ CRÉDITO	0.011253
BAJAS SEGUROS	0.003549
TINV	0.002162
INVERSIONES	0.000977
QUEJAS	6.76E-05
TOTAL	1

Tabla 16: Agrupación por fuentes de la importancia de variables para el modelo de bajas involuntarias para el mes de Enero de 2019.

Fuente	Importancia Variables Involuntarios
INGENIERÍA DE VARIABLES	0.670701
CLIENTES	0.230938
SEGUROS	0.081263
VERAZ	0.009878
VERAZ CRÉDITO	0.006161
SEGUROS ALTAS	0.000589
TINV	0.000272
SEGUROS BAJAS	0.000131
INVERSIONES	4.5E-05
QUEJAS	2.25E-05
TOTAL	1

Además, también se quiso encontrar específicamente cuáles eran las variables más importantes, por lo que se hizo una agregación por base de variables, incluyendo todas las agregaciones que se pueden haber hecho en la ingeniería de variables. Mostramos las primeras 12. Acá también normalizamos para que la suma de todas las importancias sea igual a 1.

Tabla 17: Variables más importantes para el modelo de bajas voluntarias para el mes de Enero de 2019.

Variable Base	Importancia Variables Voluntarios	Descripción de la variable
FECHA_SUSCRIP	0.080247	Meses desde fecha de suscripción de la póliza,
DIAS_EN_MORA	0.047505	Días incurridos desde que tiene mora de algún producto
HOME_BANKING_TRX	0.0403	Cantidad mensual de transacciones por home banking
MARG_PASIVO	0.03644	Margen pasivo
TIPO_CTA_A	0.028487	Cantidad de pólizas VIDA con tipo de cuenta activa
SALDO_MEDIO_CREDITO	0.024061	Saldo medio de los créditos del cliente al banco
SALDO_CUENTA1	0.023491	Saldo de Caja de ahorro en una cuenta en particular
PAS_IMPORTE	0.023085	Monto de las acreditaciones que tuvo el cliente este mes
VIDA_CUOTA	0.022856	Cuota de las pólizas de vida del cliente
SEGUROS_SUM	0.017565	Suma de la cantidad de seguros que tiene el cliente
TOTAL	0.344036	

Tabla 18: Variables más importantes para el modelo de bajas involuntarias para el mes de Enero de 2019.

Variable Base	Importancia Variables Involuntarios	Descripción de la variable
DIAS_EN_MORA	0.26981	Días incurridos desde que tiene mora de algún producto
PAS_IMPORTE	0.042956	Monto de las acreditaciones que tuvo el cliente este mes
SALDO_CUENTA1	0.042072	Saldo de Caja de ahorro en una cuenta en particular
MORA	0.029091	Marca si el cliente está en mora
TARJ_TRX_SUMA	0.025353	Agregación de la cantidad de transacciones que se hicieron con la tarjeta de crédito
TB_TARJ3_CONSUMO	0.024576	El consumo que hizo el cliente con una marca de tarjeta de crédito
FECHA_SUSCRIP	0.023977	Meses desde fecha de suscripción de la póliza
CANT_PROD_MORA	0.022663	Cantidad de productos en los que el cliente está en mora
TB_TARJ3_TRX	0.019057	Cantidad de transacciones que se hicieron una marca en particular de tarjetas de crédito
MARG_PASIVO	0.017021	Margen pasivo
TOTAL	0.516575	

En primer lugar, se muestran en las Tablas 15 y 16 que la ingeniería de variables tiene un efecto fundamental en la *performance* de los modelos. Se muestra que, sobretodo en modelos de árboles, realizar agregaciones de variables cuando se tienen resultados históricos es muy importante a la hora de realizar modelados. En ambos casos (modelos de bajas voluntarias e

involuntarias), la importancia normalizada de estas ingenierías es alrededor del triple que la importancia de la fuente más relevante en cada modelo (el modelo de bajas voluntarias tiene una importancia normalizada de 0.62 contra la importancia de 0.21 de la variable SEGUROS, mientras que el de bajas involuntarias tiene un valor de 0.67 contra la importancia de la fuente CLIENTES, con valor 0.23).

Se ven en estas tablas que en el modelo de bajas voluntarias, la fuente más importante es la de SEGUROS, mientras que en la de bajas involuntarias es la fuente de CLIENTES. En la Tabla 18 se muestran que las variables más importantes fueron las relacionadas con moras (DIAS_EN_MORA, MORA y CANT_PROD_MORA, con importancias normalizadas de 0.27, 0.03 y 0.02 respectivamente) aspecto por el cual muchos clientes pueden darse de baja involuntariamente. Es por eso que el modelo de bajas involuntarias tiene mucho mayor poder de predicción que el modelo de bajas voluntarias, los clientes que tienen cuentas pendientes con la aseguradora son más propensos a dejar de contar el servicio el próximo mes.

Por otro lado, variables que tienen que ver con las fechas de suscripción de la póliza son importantes a la hora de predecir las bajas. La variable FECHA_SUSCRIPCION en el modelo de bajas voluntarias tiene una importancia normalizada de 0.08, casi el doble que la variable siguiente que también es DIAS_EN_MORA, con 0.04. En este modelo la variable DIAS_EN_MORA es fundamental para la predicción de las bajas, pero no tanto como en el modelo de bajas involuntarias, en donde la importancia normalizada tiene un valor igual a 0.2681.

3.3 - Ganancia y respuesta

Otra forma de ver de qué manera y qué tan efectivamente se predicen las bajas es usando gráficos de ganancia y respuesta (Tsiptsis, K. K., & Chorianopoulos, A., 2011). Lo que muestran estos gráficos es: si se ordenaran los clientes de mayor a menor de acuerdo a su probabilidad predicha de churn, a qué porcentaje estaría captando el modelo.

Ganancia: De los clientes que hay hasta el momento, qué porcentaje representa la cantidad de bajas que hay hasta ahí con respecto al total de bajas. Un modelo aleatorio tendría una recta de (0,0) a (100,100).

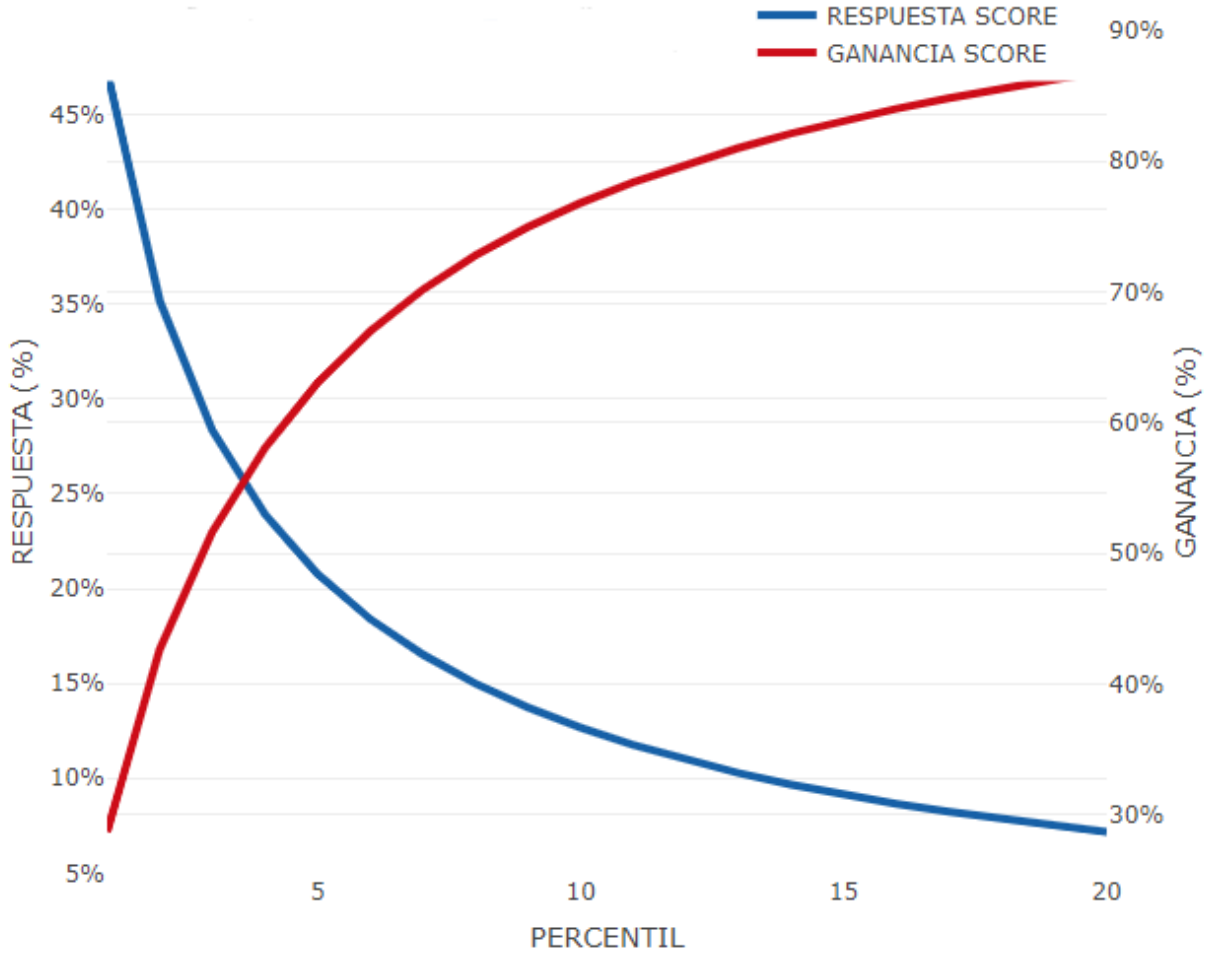
Respuesta: Qué porcentaje de los clientes que hay hasta ese percentil se dieron de baja. Si el modelo fuera aleatorio la función sería aproximadamente constante en porcentaje de bajas.

Las figuras 12 y 13 corresponden al último mes con datos recibidos.

Vemos en la Figura 12, en el modelo para bajas involuntarias, que si ordenamos a los clientes de acuerdo a su probabilidad de churn, y solamente nos quedamos con el percentil más alto, la respuesta es de 45 %. Esto significa que de los n/100 clientes más propensos a darse de baja según el modelo, el 45% efectivamente se dio de baja.

Con respecto a la ganancia, notamos que para el percentil 20, la ganancia ya se encuentra en 90%. Significa que los n/5 clientes con más propensión a churn según el modelo, son el 90% del total de clientes que efectivamente se dieron de baja.

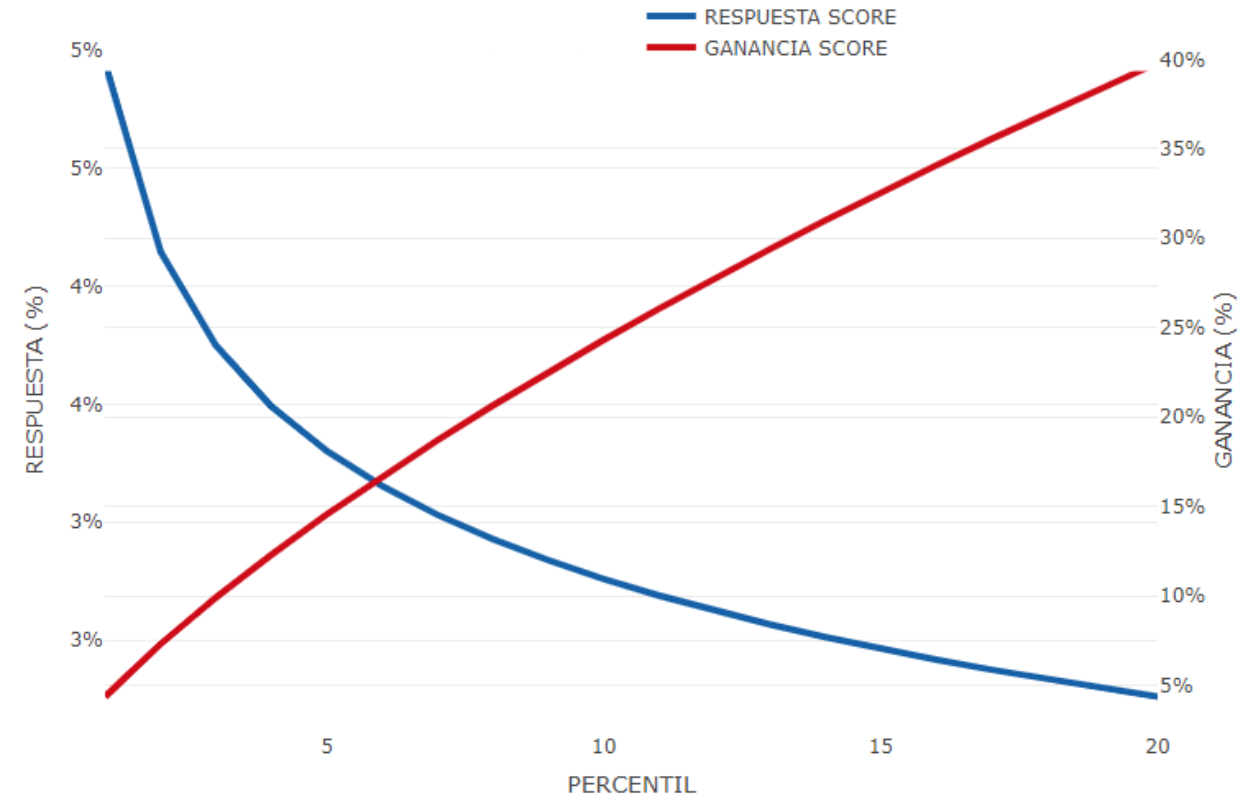
Figura 12: Gráfico de Ganancia y Respuesta para el modelo de bajas involuntarias del mes de Enero de 2019.



Para el modelo de voluntarios, vemos resultados no tan rotundos como en el modelo de involuntarios. Observamos en el percentil 1 de la Figura 13 que la respuesta es de 6,10%, lo que nos demuestra que la performance en voluntarios no es tan buena. Solamente el 6,10% de los clientes que están marcados como los más propensos a darse de baja lo hacen.

Por otro lado, vemos que la pendiente de la curva de ganancias es cercana a ser lineal, no es tan cóncava como la anterior, otro indicio que indicaría que el modelo no es tan efectivo. Igualmente, se nota que en el percentil 20, el porcentaje de ganancia se encuentra en 40%. Esto significa que si la gente de la aseguradora tuviera que elegir al 20% de la cartera con este modelo, encontraría al 40% de las personas que efectivamente realizan churn.

Figura 13: Gráfico de Ganancia y Respuesta para el modelo de bajas voluntarias del mes de Enero de 2019.



4 - Conclusiones

En el trabajo presentado se muestra cómo influyen los modelos de machine learning en tratar de evitar el customer churn y retener más eficazmente a los clientes de la compañía aseguradora. Los resultados muestran que al elegir una proporción de los clientes más propensos a darse de baja, efectivamente una gran parte de estos lo hace. Por lo cual se podrían empezar a identificar a estos clientes para poder realizar alguna acción al respecto, como promociones o campañas para evitar que realicen churn.

Estos resultados ayudan a que la compañía se concentre mejor en los usuarios con más probabilidades de dejar de contar con sus servicios. Esto permitiría ahorrar una buena cantidad de cuotas futuras en ellos, es decir evitar hacer campañas con clientes al azar, para no desperdiciar tiempo y mano de obra en los consumidores que sin importar la llamada seguirán comprando el servicio.

Por cuanto a las campañas que se realizan para evitar el churn, existen algunas cuestiones futuras a tener en cuenta. En primer lugar, si se empiezan a considerar acciones en donde se llamen a los clientes con mayor probabilidad de realizar churn, los modelos comenzarían a fallar dependiendo de qué tan efectivas sean las campañas. Por lo tanto es importante hacer un buen seguimiento de los estados de las campañas hacia los clientes, y en lo posible tratar de averiguar si hubo *compliance*, es decir que el cliente haya contestado la llamada, mail, SMS o la vía por la cual se lo haya contactado, y haya recibido el mensaje de la campaña, así como también el momento preciso en el que lo recibió.

Para la implementación, se podría empezar a investigar sobre qué porcentaje de los mejores puntuados se empezaría a actuar. Llamar al 1% mejor puntuado de los clientes indicaría una buena tasa de respuesta, ya que la mayoría de estos clientes se darían de baja si no se aplicaran estas campañas hacia ellos, pero se podrían llamar más clientes para disuadirlos de realizar el churn. Por otro lado, si se llaman a muchos clientes para tratar de erradicar este problema, se generan más costos de mano de obra para la gente de tele-marketing, en donde no

pueden generar tantos retornos, ya que los rendimientos son decrecientes (si se quiere llamar a un cliente más, la probabilidad de compra predicha por el modelo será indefectiblemente menor o igual al inmediato anterior). Un buen equilibrio generaría las mejores ganancias, por lo tanto un análisis de costos y beneficios vendría bien como trabajo futuro a analizar, de acuerdo a las probabilidades que brinda el modelo.

Por otro lado, este trabajo se hizo con el modelo LightGBM, pero se podría profundizar como trabajo futuro el evaluar la performance de otros clasificadores para tratar de mejorarla, como por ejemplo redes neuronales (Au, W. H., Chan, K. C., & Yao, X., 2003; Soeini, R. A., & Rodpysh, K. V., 2012).

Además, se considera fundamental la actualización de los modelos cada mes para poder entrenarlos con los datos más actualizados posibles. De esta manera no se perdería eficacia a la hora de predecir las bajas potenciales. Y el sistema contaría con los nuevos clientes y permitiría un mayor alcance.

Otra cosa importante es tratar de entender cuál es el *target* a llamar de la aseguradora. Probablemente, si un cliente tiene una probabilidad muy alta de darse de baja, una llamada de la gente de telemarketing no impedirá que el sujeto realice churn. Para esto es importante contar con diferentes tipos de estrategias para buscar la permanencia del cliente en la aseguradora. El modelo permitirá identificar cuáles son las características principales de los clientes para poder enfocar en determinadas cualidades las estrategias de convencimiento para que permanezcan en la firma. (Kuhn, M., & Johnson, K. ,2013).

En la aseguradora, algunos clientes pueden ser gestionados para tratar de evitar su deserción, mientras que otros no serán contactados. Existen escenarios en donde los clientes gestionados cambiarán de opinión, y otros casos donde su elección no podrá ser cambiada por una gestión. El objetivo a largo plazo es evitar la menor cantidad de bajas posibles.

Podemos identificar 4 tipos de casos, indicados en la tabla 18.

Tabla 18: Escenarios de tipos de clientes de acuerdo a su acción dependiendo si son gestionados o no.

No gestionado	**Gestionado**	
	Realiza Churn	**No realiza churn**
Realiza churn	Realiza churn independientemente de si fue gestionado o no.	**Cliente que cambia de opinión por el hecho de estar gestionado.**
No realiza churn	Cliente que tienen una respuesta negativa a la gestión y realizan churn	El estar o no gestionado no afecta, el cliente no se dará de baja.

La verdadera búsqueda consiste en encontrar a los clientes que cambian de parecer por el hecho de estar gestionados, que no necesariamente es la gente que más probabilidad de churn tiene. Es por eso que del dicho al hecho puede haber inconsistencias. El modelo identifica los clientes más cercanos a darse de baja pero no toma en cuenta aquellos que ya están decididos a abandonar el servicio y no cambiarán de parecer.

Por las probabilidades predichas en este trabajo generalmente los clientes no tienen probabilidades muy altas de hacer churn. Sin embargo, a medida que el modelo aumenta su grado de precisión, la importancia del modelo quedará expuesta y empezará a quedar obsoleto y perderá valor.

Para posteriores trabajos se pueden incluir detallados análisis de la elección óptima de la ventana de tiempo, junto con las acciones de las campañas. También se podrían realizar experimentos de cuáles son las campañas más efectivas a la hora de retener clientes, y cuánto dinero se ahorra la aseguradora por poner en marcha estas investigaciones.

5 - Bibliografía

- Au, W. H., Chan, K. C., & Yao, X. (2003). A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE transactions on evolutionary computation*, *7*(6), 532-545.
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*(Vol. 26). New York: Springer, 522,523.
- Japkowicz, N. (2003, August). Class imbalances: are we focusing on the right issue. In *Workshop on Learning from Imbalanced Data Sets II* (Vol. 1723, p. 63).
- Visa, S., & Ralescu, A. (2005, April). Issues in mining imbalanced data sets-a review paper. In *Proceedings of the sixteen midwest artificial intelligence and cognitive science conference* (Vol. 2005, pp. 67-73). sn.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (pp. 3146-3154).
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.
- Kotler, P., Keller, K. L., Brady, M., Goodman, M., & Hansen, T. (2016). *Marketing management*. Pearson Education Ltd..
- Staudt, M., Kietz, J. U., & Reimer, U. (1998, August). A Data Mining Support Environment and its Application on Insurance Data. In *KDD* (pp. 105-111).
- Domeniconi, C., Perng, C. S., Vilalta, R., & Ma, S. (2002, August). A classification approach for prediction of target events in temporal sequences. In *European Conference on*

Principles of Data Mining and Knowledge Discovery (pp. 125-137). Springer, Berlin, Heidelberg.

- Guillen, M., Nielsen, J. P., & Pérez-Marín, A. M. (2008). The need to monitor customer loyalty and business risk in the European insurance industry. *The Geneva Papers on Risk and Insurance-Issues and Practice*, *33*(2), 207-218.

- Guillén, M., Nielsen, J. P., Scheike, T. H., & Pérez-Marín, A. M. (2012). Time-varying effects in the analysis of customer loyalty: A case study in insurance. *Expert systems with Applications*, *39*(3), 3551-3558.

- Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of marketing research*, *43*(2), 204-211.

- Berry, M. J., & Linoff, G. S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.

- Tsiptsis, K. K., & Chorianopoulos, A. (2011). *Data mining techniques in CRM: inside customer segmentation*. John Wiley & Sons.

- Soeini, R. A., & Rodpysh, K. V. (2012). Applying data mining to insurance customer churn management. *Int. Proc. Comput. Sci. Inf. Technol*, *30*, 82-92.

- Günther, C. C., Tvete, I. F., Aas, K., Sandnes, G. I., & Borgan, Ø. (2014). Modelling and predicting customer churn from an insurance company. *Scandinavian Actuarial Journal*, *2014*(1), 58-71.

- Chen, T., & He, T. (2015, August). Higgs boson discovery with boosted trees. In NIPS 2014 workshop on high-energy physics and machine learning (pp. 69-80).