Prof. Jaime Miranda y Felipe Bugueño

APLICACIONES DE INTELIGENCIA DE NEGOCIOS

CURSO BUSINESS INTELLIGENCE

ENGIN460/01 | TAREA 2

ANÁLISIS DE MODELOS PREDICTIVOS Y OPTIMIZACIÓN DE RESULTADOS

FACULTAD ECONOMÍA Y NEGOCIOS UNIVERSIDAD DE CHILE

Integrantes

Camila Bustos
Simón Durán
Luis Marilao

Resumen ejecutivo

El problema analizado en el presente documento contempla la decisión de una compañía aseguradora de si debe o no llevar a análisis las solicitudes de cobro de sus asegurados. Esto en concordancia con generar un máximo beneficio, entregado al maximizar el ahorro total mientras se minimiza el costo de concretar los análisis.

En principio, se introducen las variables capturadas por la compañía de cada solicitud en un modelo predictivo, donde se realiza un trabajo de comparación entre una serie de modelos y combinaciones de factores como cantidad de datos, variables, etc. con el objetivo de generar la mejor predicción de cuáles de esas instancias corresponden a posibles fraudes por parte de los asegurados.

Hipótesis: *La compañía genera un beneficio significativo al detectar solicitudes fraudulentas, aun teniendo que incurrir en costos para confirmar su falsedad.*

Para el estudio de esta hipótesis, se revisa un problema de optimización que, mediante la asignación de casos que la compañía debe revisar, la determinación de la cantidad de abogados a contratar y el recuento de cuántos casos analizar para cada tipo de seguro que gestiona esta empresa, logra encontrar la combinación de elementos que lleva a la compañía a tener un máximo beneficio monetario.

Como resultado de incorporar la predicción en el modelo de optimización, se obtiene que la compañía puede obtener un beneficio de $125.131.362, a costa de evitar pagar dinero solicitado por clientes y llevando sus casos particulares a análisis. Además, se desprende que siempre la compañía enviará a todos sus abogados a disposición a revisar instancias y que los tipos de seguros que están con mayor frecuencia en la palestra de análisis son los seguros del tipo *Enfermedad (i=1)* y *despido (i=5)*, los cuales se estiman como los más caros de pagar, razón por la que generalmente es más conveniente llevarlos a análisis, junto a ser las causas con en las que se presenta mayor proporción de solicitudes fraudulentas.

Se pone en cuestión la resolución general de la compañía debido a que se carece de información relativa a lo que sucede luego de haber analizado los casos. Dicho análisis permitirá ver el beneficio real, ya que capturará el costo de haber analizado un caso e incurrir en su pago, dado que no era un caso fraudulento, sin embargo, al no disponer de estos datos, se dejará este análisis para otro estudio del problema en el futuro.

Tabla de contenido

Introducción

La detección de fraudes en las solicitudes de cobro de seguros en una compañía de la industria aseguradora es un proceso fundamental para estas. Los desembolsos incurridos en el pago de seguros reclamados suelen ser altos, razón por la cual la compañía siempre esperará pagar la menor cantidad de dinero posible, es decir, evitar incurrir en desembolsos por casos fraudulentos cuando estos superan el costo de analizar su veracidad. Análisis de expertos que se dedican a elaborar perfiles para clasificar a clientes rentables y no rentables, o bien, clientes con baja y alta probabilidad de cobrar un seguro, generan datos que apoyan a etiquetar algunos cobros como posible fraude, donde se requiere de un análisis manual posterior para verificar cada caso detectado.

Si bien las compañías pierden dinero al pagar por cobros de casos falsos, también deben evaluar el costo de oportunidad existente entre enfocar recursos a la detección de estos y pagar algunos de ellos; el desembolso requerido frente a algunas solicitudes no es tan relevante en comparación al tiempo, esfuerzo y dinero necesarios para analizar su veracidad. Razonablemente, resulta un punto de interés en estas firmas encontrar metodologías que permitan automatizar o facilitar este proceso de análisis, donde el *data mining* y el *machine learning* pueden ser de gran apoyo.

Data mining es una de las etapas de *Knowledge Discovery in Databases* (KDD) donde, mediante las ciencias de la computación, se intentan descubrir patrones en grandes volúmenes de datos, utilizando la inteligencia artificial, el aprendizaje automático (machine learning), elementos estadísticos y sistemas de bases de datos. Estos patrones suelen no poder identificarse de forma manual o rápida, motivo por el cual este campo es particularmente útil, ya que entrega información relevante que no había sido antes descubierta, como agrupaciones de datos o clusters, anomalías, interdependencia de datos, entre otros.

En el presente documento, se recurre a las etapas de KDD, mediante modelos predictivos de data mining, con el objeto de detectar los elementos antes mencionados en un conjunto de datos disponibles sobre la solicitud de cobro de seguros en una compañía. El resultado favorable de esta metodología permite automatizar el proceso de detección de fraudes, disminuyendo la probabilidad de incurrir en desembolsos por este concepto.

Posteriormente, se analiza un problema de optimización enfocado a dicho resultado, donde se genera el óptimo de casos a analizar, intentando maximizar el beneficio en consideración al costo de pagar solicitudes versus el de analizarlas. Para su revisión se despliega un análisis de resultados y sensibilización, a partir de lo cual se obtienen lineamientos de decisión para esta compañía.

Finalmente, a partir de los procedimientos mencionados, se pretende responder a la hipótesis de que la detección de casos fraudulentos impacta significativamente en el beneficio monetario capturado por la compañía, aun cuando se incurran en costos para verificar su falsedad.

Análisis Exploratorio y Selección de Variables

En primer lugar, se analiza el comportamiento y distribución general de la clase, respecto a la cuantía y proporción cuando esta toma el valor de "POSIBLE FRAUDE" y "NO". Se extrae la tabla 1, en la que se observa una clase desbalanceada, es decir con la mayor parte de los datos pertenecientes a la clase "NO". Es por ello que en el posterior proceso de *data mining* se balancearán los datos con el objetivo de construir un modelo predictivo más preciso.

Adicionalmente, un criterio relevante de selección será si estas superan el umbral de 5,82% o 6% aprox. de posible fraude, correspondiente al porcentaje de la muestra general.

CLASE	OBS	PORCENTAJE
NO	315577	94,18%
POSIBLE FRAUDE	19497	5,82%
TOTAL	335074	100,00%

Tabla 1. Distribución de la muestra general.

Para cada variable disponible en la base de datos se realizó un análisis particular. Dado que son una gran cantidad, este análisis particular se anexará al final de este documento **(Anexo 1. Análisis particular detallado).** A continuación, se describe el análisis realizado para las variables mas relevantes en el modelo, con el fin de generar un perfil de instancias fraudulentas.

Primero se eliminan las variables SIN_ID, LOT_NUMEROLOTE, NAME, LASTNAME, ID_PRODUCTO_TECNICO, EVA_ID, CUO_NUMERO_CUOTA, NAME_DENUNCIANTE y RAMO_LEGAL. Esto es dado que son, en su mayoría, atributos que identifica a cada una de las instancias en distintos ámbitos, es decir, no poseen un impacto en la predicción de fraudes, sino que solo sirve para identificar a las instancias. Además, no se encuentran relaciones causales entre ellas, sino solo relaciones espurias, por ejemplo, no se espera que el nombre de una persona determine si esta comete fraudes o no. Se desconoce a que hace referencia la variable CUO_NUMERO_CUOTA, sin embargo, a pesar de ello, sus filas tienen solo valores NULL (vacíos) motivo por el cual se decide eliminar la variable del modelo. En suma, se decide eliminar estas variables por irrelevancia y falta de valor predictivo.

- **TIPO_EMPRESA_SEGURO:** Representa la industria a la cual pertenece la empresa asegurada que solicita el cobro. Su distribución general es:

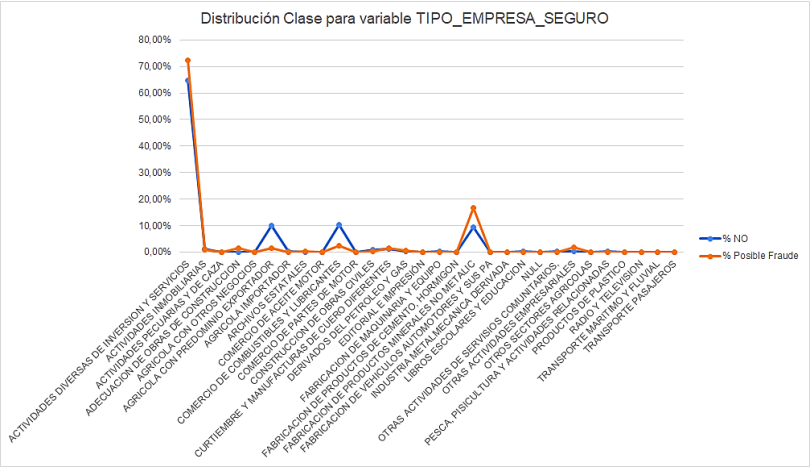


Gráfico 1. Distribución de probabilidad de TIPO_EMPRESA_SEGURO respecto a valores de CLASE.

Esta variable es importante para el modelo predictivo ya que entrega valores que tienen una tasa de posibles fraudes significativas en cantidad, distribución de probabilidad y porcentaje de posible fraude. Se genera un gráfico resumen de las distribuciones, las cuales posteriormente se harán binarias para el modelo predictivo. Adicionalmente, se eliminan 7 filas que toman valores NULL, principalmente porque no aportan para predecir un posible fraude.

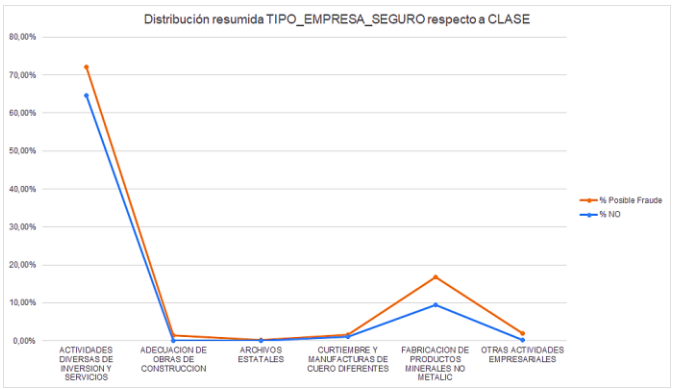


Gráfico 2. Distribución de probabilidad de TIPO_EMPRESA_SEGURO (valores con mayor % Posible Fraude) respecto a valores de CLASE.

- **FECHA_NACIMIENTO:** De esta variable se deriva una nueva variable, la de EDAD, donde se grafica su comportamiento respecto a los valores de CLASE.

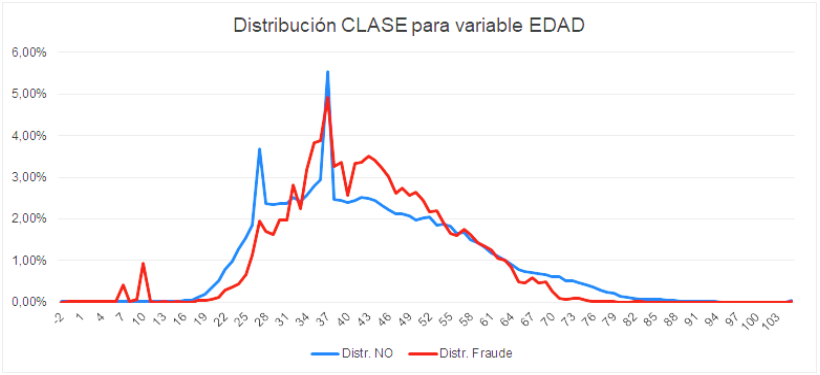


Gráfico 3. Distr. De clase para variable Edad.

Se puede ver que hay un rango etario que concentra una mayor tasa de fraude, este es de los 35 a 50 años aproximadamente. Además, se eliminan las filas que contengan edades negativas, ya que no sigue la lógica de la edad en números naturales, además, no hay una implicancia que dichas edades conduzcan a posible fraude, más bien sigue la misma distribución que la base de datos misma.

Se identifican dos tuplas con esta variable vacía y de clase no, para los SIN_ID 3246369 y 3472994 que son eliminadas desde Access.

- **BANCO:** Respecto a la variable BANCO, se realizó un gráfico de distribución de probabilidades sujeto a los valores de la variable CLASE. Los valores ABB, RCO y SCC son indicadores de posibles fraudes a modo visual; y AFA, MFO, NBA y OBC también indicadores de posibles fraudes con un ratio mayor al 7% de fraudes en las instancias totales para cada uno.

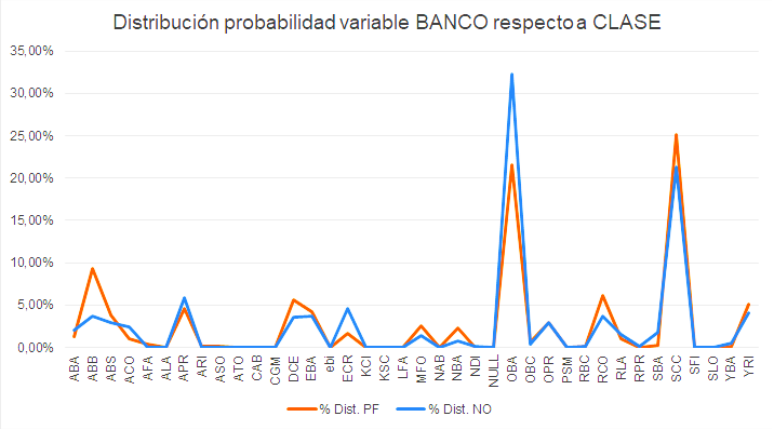


Gráfico 4. Distribución de probabilidad variable Bco respecto a clase.

Se borran 11 filas que corresponden a valor NULL que tienen CLASE igual no.

- **USER_LIQUIDADOR:** La variable representa quien gestionó la venta del seguro. Se observa que un 70% aprox. de la muestra fue gestionado por una empresa externa (40% EmpresaExterna1 y 33% EmpresaExterna1Digital). En suma, de los casos catalogados como posibles fraudes, un 52% fueron gestionados por empresa externa 1 y un 20% EmpresaExterna1 Digital (72% total), visualmente:

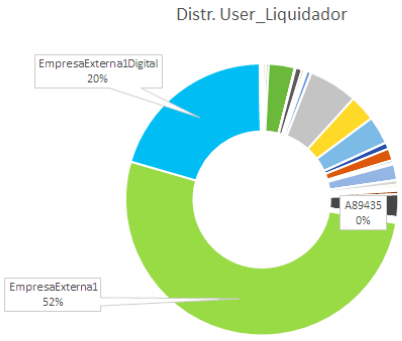


Gráfico 5. Distr USER_LIQUIDADOR

Sin embargo, al realizar un análisis particular para cada una de las entidades que gestionan los seguros (empresas externas e internas) y evaluar cuánto representan los casos de posibles fraudes respecto a sí mismas, se extrae que la EmpresaExterna1 posee un 8% de fraudes, EmpresaExterna1Digital un 4% y la entidad 530815 un 9%.

Esto implica que, al estar sobre el porcentaje esperado de posibles fraudes evaluado en la base general, solo las entidades EmpresaExterna1 y 530815 se quedarán dentro del modelo. Ante ello, se

generan 2 nuevas variables binarias: (1) Emp1 que toma valor 1 si pertenece a esa empresa y 0 si no, y (2) Emp2 que toma valor 1 si corresponde a la entidad 530815 y 0 si no.

- SIN_NOMBRE_CIUDAD: Representa la ciudad de residencia de la persona que solicita el cobro de seguro. Observamos primero las ciudades que tienen más casos de posible fraude respecto al total de solicitudes de cobro ocurridas en la misma ciudad. Las primeras 5 ciudades ordenadas por cómo se describe, se muestran en la siguiente tabla:

CIUDAD	POSIBLE FRAUDE	NO	Total	%PF x Ciudad	%PF x total PF
San Gregorio	3	2	5	60%	0%
Pirque	1	3	4	25%	0%
Quellón (Puerto Quellón)	1	4	5	20%	0%
Licantén	7	36	43	16%	0%
Antuco	10	56	66	15%	0%

Tabla 2. Distr. ciudades por Clase

Notar que, aun cuando un 60% de las solicitudes de San Gregorio fueron posibles fraudes, son solo 5 los casos de solicitud en esta ciudad, siendo muy poco significativa en la muestra. Por este motivo, se observan cuántos casos detectados como posibles fraudes respecto al total de posibles fraudes detectados, hay en cada ciudad:

CIUDAD	POSIBLE FRAUDE	NO	Total	%PF x Ciudad	%PF x total PF
Santiago	8284	134920	143204	6%	42%
Rancagua	662	6609	7271	9%	3%
Antofagasta	534	7978	8512	6%	3%
Concepción	517	4427	4944	10%	3%
La Serena	421	5030	5451	8%	2%

Tabla 3. Ciudades ordenadas

Santiago representa un 42% de los casos posibles fraudes, pero a su vez, se estima que puede deberse a que esta ciudad representa un alto porcentaje de solicitudes generales respecto a la muestra (Coincidentemente, un 42% también). La ciudad de Santiago será parte del perfil de posibles fraudes. Para su inclusión en el modelo, se creará una binaria "Stgo" que tomará el valor 1 si el caso pertenece a Santiago y 0 si no.

- CAU_DESCRIPCION: En primera instancia se observa que la causa de solicitud más común en la muestra es Despido, representando un 65% de las causas. Esto genera un sesgo a priori, donde solo por ser mayoritaria en la muestra genera un alto porcentaje de presentar casos de posibles fraudes. Se analiza entonces la cantidad de casos detectados como posible fraude para cada causa de solicitud, donde se observa que aquellas que presentan un impacto más relevante son las destacadas en amarillo que se muestran a continuación:

Causa Solicitud	Fraudes	Total	Fraudes/total
Despido	14104	218669	6%
Enfermedad	3467	35809	10%
Defunción	261	31303	1%
Fraude	307	21224	1%

Clonación	213	12835	2%
NULL	179	4397	4%
Robo	453	2921	16%
Enfermedades Graves	56	2303	2%
Diagnóstico	100	2105	5%
Reembolso Gastos Médicos	19	1954	1%
Incendio	16	571	3%
Siniestro Pagado E.G.	204	322	63%
Sismo	4	204	2%
Accidente	7	180	4%
Desastre Natural / Temporal	27	151	18%
otra	59	59	100%
Protección Patrimonial	1	28	4%
Daños por terceros	3	17	18%
Siniestro Pendiente E.G.	14	14	100%
Inundación	3	7	43%
Siniestros Pagados Salud Caja Los Andes	0	1	0%
Total general	**19497**	**335074**	**6%**

Tabla 4. Causa solicitud de cobro.

Los criterios para seleccionar qué causas se quedarán en el modelo predictivo final, son (1) El volumen de la causa respecto al total de instancias, ya que debe ser representativo en la muestra. Por ejemplo, si bien en la causa "enfermedad" solo un 10% son casos de posible fraude, en general esa variable representa la segunda mayoría respecto de la muestra total, siendo un 11% del total de la muestra. (2) El nivel de casos detectados como posible fraude respecto al total de casos para cada una de las causas. Este factor es relevante porque hay tipos de causas que su 100% (o un alto porcentaje) son posibles fraudes, por ejemplo, el 63% de las solicitudes bajo la causa "Sinestro pagado E.G" son casos de posible fraude. Esto implica que, si una solicitud se hace bajo ese motivo, tendría una alta probabilidad de ser una solicitud fraudulenta.

Finalmente, se binarizarán las 5 causas escogidas y se eliminará la variable original.

Análisis de relación entre variables

Al estudiar la base de datos se observan algunas incongruencias:

- Para los casos de 'Defunción', hay aproximadamente 19.000 instancias que tienen un DNI_ASEGURADO igual al DNI_DENUNCIANTE, que hace poco sentido ya que un fallecido no puede ir a denunciar su propia defunción. Se realizó un análisis respecto a la variable CLASE pero para el total de casos mencionados, menos de un 0,1% pertenecía a POSIBLE FRAUDE, por lo que no se consideró realizar una nueva variable respecto a esta incongruencia para ser agregada al modelo de predicción.

- Se analiza cuando un siniestro pertenece a un menor de edad y se da cuenta que hay instancia que corresponden a CAU_DESCRIPCIÓN 'Despido' lo que no hace sentido ya que solo mayores de 18 años pueden trabajar con contrato, y además, no se debiera analizar

seguros de despido para estos. Para estos casos se eliminaron dichas instancias (249). En tanto a su comportamiento, solo un 0,2% pertenecía a Posible Fraude.

Variables derivadas:

- Se realizó la diferencia de días entre SIN_FECHA_OCURRENCIA y SIN_FECHA_DENUNCIA para ver la relación entre la demora de denunciar un siniestro desde su ocurrencia y los 'Posibles Fraudes'. A continuación, el gráfico:

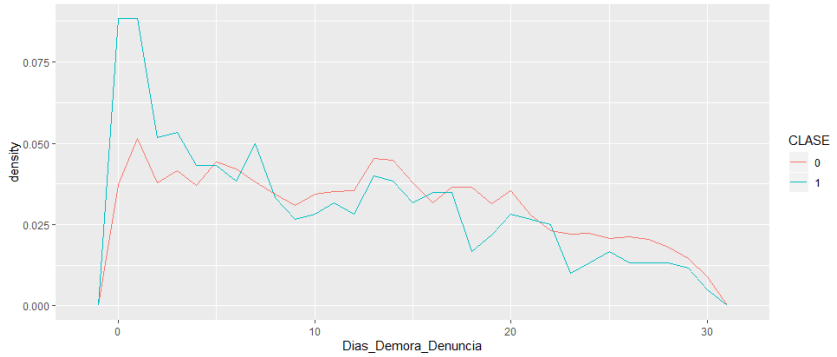


Gráfico 6. Distribución de los días que se demora en denunciar un siniestro respecto a los valores de CLASE.

Se puede apreciar que hay una concentración de 'Posibles Fraudes' cuando la denuncia se realiza antes de los 5 días después de ocurrido el siniestro. De esto, se crea la variable binaria Demora_Denuncia_Siniestro que toma valor 1 cuando los días de demora en denunciar es igual o menor a 5.

- Se creó una variable que considera la diferencia entre HEV_TIME_MARK y LOT_FECHA_RECEPCION para obtener información sobre los días de demora que hay entre la fecha de llegada de un siniestro a la compañía y su correspondiente gestión, y cómo se relaciona con los 'Posibles Fraudes': A continuación, el gráfico que muestra el comportamiento:

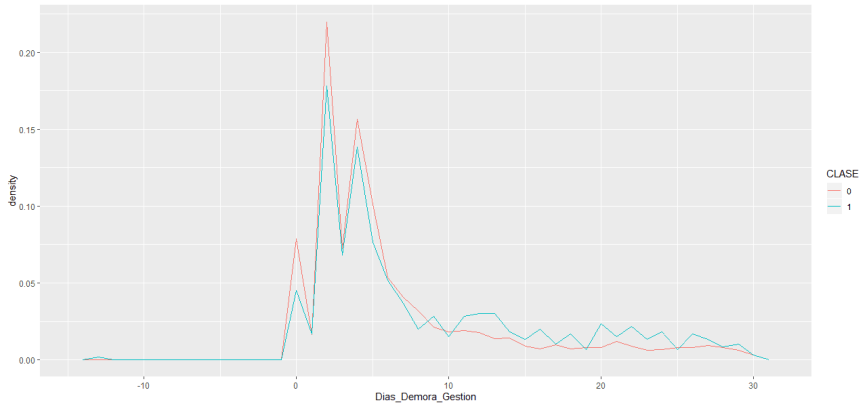


Gráfico 7. Distribución de los días de demora de gestión respecto a los valores de CLASE.

Del gráfico anterior se puede rescatar que hay tasas de posible fraude mayores a no cuando la demora en gestionar un siniestro es mayor a los 10 días y menor a 30 días. Dado esto, se crea una variable binaria que captura este comportamiento, tomando valor 1 cuando los días de demora en gestionar se encuentra entre 10 y 30 días.

- De la base de datos Siniestros, se realizó una suma de los montos aprobados históricos para cada DNI_ASEGURADO para que sirvan como antecedente y características de dichos asegurados. Además, esta nueva variable 'Suma_Monto' fue agregada a la base de predicción, ya que en aquella base estaban registrados DNI_ASEGURADO que ya se contenían en base Siniestros. Solo 2 instancias de la base predicción no tenían 'Suma_Monto' (NULL) dado que no tenían historial de montos aprobados en la base histórica de siniestros, por lo que se asoció un valor 0 a esta variable para dichas instanciaS, dando a interpretar que son DNI_ASEGURADO que nunca han realizado cobros o reclamos de seguros, por ende, nunca se la ha aprobado montos.

- De la base de devolución se rescató la variable POLIZA_SEGURO con el fin de rescatar las pólizas con más devoluciones y sumarlas al modelo predictivo. A continuación, la tabla con la cantidad de devoluciones por póliza.

POLIZA_SEGURO	Cantidad Devoluciones
58013392	128144
108050564	81385
56567116	59128
58053530	43895
108510574	30606
57032664	25622
10709442	19506
108525576	16691
57558342	15697
56567115	15624

Tabla 5. Cantidad de devoluciones para las 10 pólizas con mayor devolución.

Dado lo anterior, se crea una variable binaria que toma valor 1 cuando una instancia toma los valores de POLIZA_SEGURO de las 3 pólizas con más devoluciones.

Finalmente, las variables utilizadas como base para el modelo predictivo se sintetizan en la tabla como sigue:

	Variable	Descripción	Naturaleza
1	LOT_FECHA_RECEPCION	Fecha escalada de cuando se recibe la solicitud de cobro	Real entre 0 y 1
2	POLIZA_SEGURO	Identificador de la póliza del seguro que se está solicitando	Factor
3	PS_XX [XX=Numero póliza]	Pertenencia a un tipo de póliza particular. Fue binarizada respecto a Póliza Seguro solo para aquellas pólizas que contenían mayor índice de fraudes. [Binarizadas: 52, 53, 57 Y 60]	Binaria
4	SIN_FEC_DENUNCIA	Fecha en que se denuncia el siniestro	Real entre 0 y 1
5	SIN_FEC_INIC_VIG_PRODUCTO	Fecha escalada cuando se inicia la vigilancia de un producto	Real entre 0 y 1
6	EDAD	Edad del asegurado, derivada de variable FECHA_NACIMIENTO. Variable escalada entre 0 y 1	Real entre 0 y 1
7	EDAD_35_50	Clasificación de si el asegurado se encuentra en un rango etario entre 35 y 50 años (Edad con mayor probabilidad de fraude)	Binaria.

8	BANCO	Código institución bancaria intermediaria en el cobro de un seguro.	Factor
9	BANCO_XX [XX=Código banco]	Clasificación de pertenencia a banco, solo para los que presentaron mayor índice de fraudes. [Bancos: ABB, RCO, SCC, AFA, MFO, NBA y OCB]	Binaria.
10	EMPXX [XX=n°empresa]	Clasificación de si la empresa XX gestionó el seguro. Empresas analizadas son las que presentan mayor índice de fraudes [EmpresaExterna1 y 530815]	Binaria
11	HD_Dictamen_Aprobado	Identificador de HEV_DETAILS cuando toma el valor Dictamen Aprobado	Binaria
12	MONTO_APROBADO	Monto (Dinero) que se pagará a los asegurados	Entero
13	SNC_STGO	Clasificación de si el siniestro pertenece a Santiago, ya que es la ciudad con mayor índice de fraudes	Binaria
14	TES_XX [XX= Tipo empresa seguro]	Clasificación de si siniestro proviene de uno de los 5 Tipos de Empresas con mayor índice de fraudes. (5 variables binarias)	Binaria
15	SIN_FECHA_OCURRENCIA	Fecha escalada en que ocurre el siniestro.	Real entre 0 y 1
16	PRODUCTO_XX [XX=Primeros 2 dígitos de n° producto]	Clasificación de si pertenece a las líneas de producto más fraudulentas, análisis obtenido desde la categorización de las líneas de productos utilizando los primeros dos dígitos de cada número de producto.	Binaria
17	DNI_ASEGURADO_RUTXX [XX=Rut asegurado]	Identificador de asegurado. Binarizada desde DNI_ASEGURADO para aquellos DNI que contenían mayor índice de fraudes. [Binarizados: 10900762512, 10900122212, 109K0135012]	Binaria
18	SIN_CUOTAS_COBERTURA	Cantidad de cuotas que cubre el seguro contratado.	Real entre 0 y 1
19	SSC_1_6	Clasificación de pertenencia a SIN_CUOTAS_COBERTURA. Binarizada para rango de cuotas entre 1 y 6 que son las que contenían mayor porcentaje de posible fraude.	Binaria
20	SIN_FEC_INIC_VIG_SINIESTRADO	Fecha escalada en que se contrata el seguro	Real entre 0 y 1
21	HEV_TIME_MARK	Fecha escalada en que se gestiona el siniestro en la compañía de seguros.	Real entre 0 y 1
22	SIN_ESTADO_ACTUAL	Estado en el que se encuentra el proceso de cobro del siniestro	Factor
23	LOT_EXTERNO	Clasificación de si las gestiones fueron realizadas por una empresa externa o no.	Binaria
24	CAU_DESCRIPCION	Causa por la cual se solicita el cobro del seguro	Factor
25	CD_XX	Identificador de valores de CAU_DESCRIPCION. Binarizada para descripciones con mayor porcentaje de posible fraude. [Binarizada Despido,Enfermedad, Robo, Siniestro_Pagado,]	Binaria
26	SUMA	Suma de monto aprobado para cada asegurado posteriormente escalada	Real entre 0 y 1
27	CLASE	Clasificación de la instancia respecto a si es fraude o no	Factor
28	Días_Demora_Denuncia	Diferencia en días entre SIN_FECHA_OCURRENCIA Y SIN_FECHA_DENUNCIA	Entero
29	Demora_Denuncia	Clasificación si Días_Demora_Denuncia toma valor menor a 5 o no	Binaria
30	Días_Demora_Gestión	Diferencia en días entre HEV_TIME_MARK y LOT_FECHA_RECEPCION	Entero
31	Demora_Gestion	Clasificación si Días_Demora_Gestión toma valor mayor a 10	Binaria
32	PS_Mas_Devolución	Clasificación si POLIZA_SEGURO toma los valores "58013392", "108050564" o "56567116"	Binaria
33	PS_Mas_Fraudulentas	Clasificación si POLIZA_SEGURO toma los valores "58068527", "5205583", "57543237", "5355687", "57017465"	Binaria

Tabla 6. Síntesis variables base para ejecución de modelos predictivos.

Sin perjuicio de que se realizaron pruebas de modelos incluyendo variables adicionales y/o quitando algunas de las que en la lista aparecen.

Preguntas de análisis

i) Grado de relación entre fraudes y reclamos

En primer lugar, se procede a vincular las bases de datos SiniestrosBD y Reclamos, mediante DNI_ASEGURADO, donde podemos ver a una sola tupla y, por tanto, un solo asegurado con DNI 10980188712, que se encuentra en ambas bases de datos, pero siempre (18 instancias) ha estado asociado a la CLASE NO. De este modo buscamos nuevas formas de vincular ambas bases para buscar una relación mediante otra variable.

No se observaron relaciones por la variable "No Column Name" porque se desconoce a qué hace referencia. Se observa que solo es posible vincular ambas bases por medio de DNI_ASEGURADO, POLIZA_SEGURO, PRODUCTO Y BANCO.

Al vincular por POLIZA_SEGURO, se identifican solo 673 distintas en base SiniestrosBD y en la base Reclamos 560 distintos. Al realizar el cruce se extrae que hay 346 POLIZA_SEGURO que están en ambas bases de datos.

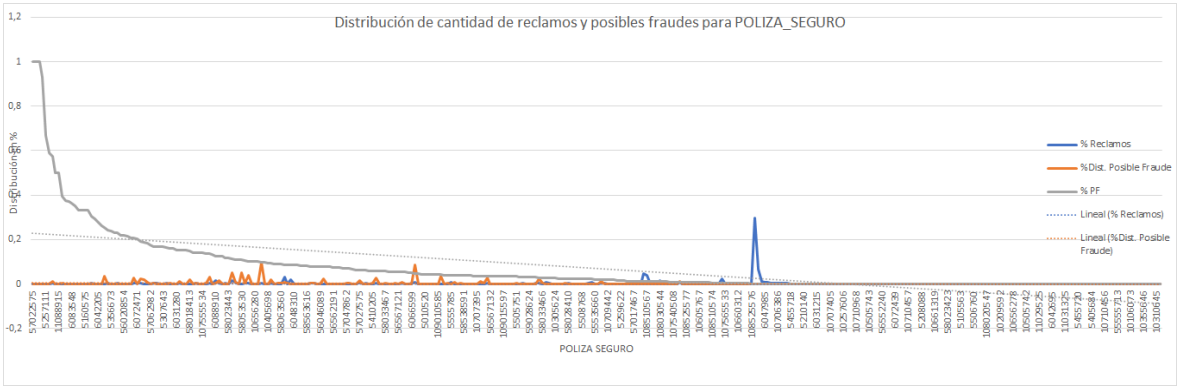


Gráfico 8. Distribución de reclamos, distribución de posibles fraudes y tasa de fraudes para las pólizas encontradas en base de reclamos y siniestro.

Del gráfico anterior, se puede comprender que la tasa de reclamos de una póliza no dicta relación con las pólizas con tasas de fraude elevadas o pólizas con mayores en distribución de posibles fraudes. De hecho, al realizar la línea tendencial de los reclamos, se mantiene con una pendiente cercana a 0, siendo que la gráfica está representada por el orden de pólizas de la más a menos fraudulenta. Por último, el coeficiente de correlación para % Reclamos y % Distribución de posibles fraudes es de 0.02 y % Reclamos con % PF es de -0.03, es decir, una relación muy cercana a 0.

Vinculando por PRODUCTO, primero se observa que hay 486 PRODUCTOS distintos en la base Reclamos, y 630 ID_PRODUCTO_TECNICO distintos en la base SiniestrosBD. Al cruzarlos, se extraen solo 345 ID_PRODUCTO_TECNICO o PRODUCTO distintos en ambas bases, pero representan 303.353 filas. De ello, se identifica que de las personas que reclaman su ID_PRODUCTO_TECNICO o PRODUCTO no es una variable que muestre una mayor relación con POSIBLE FRAUDE.

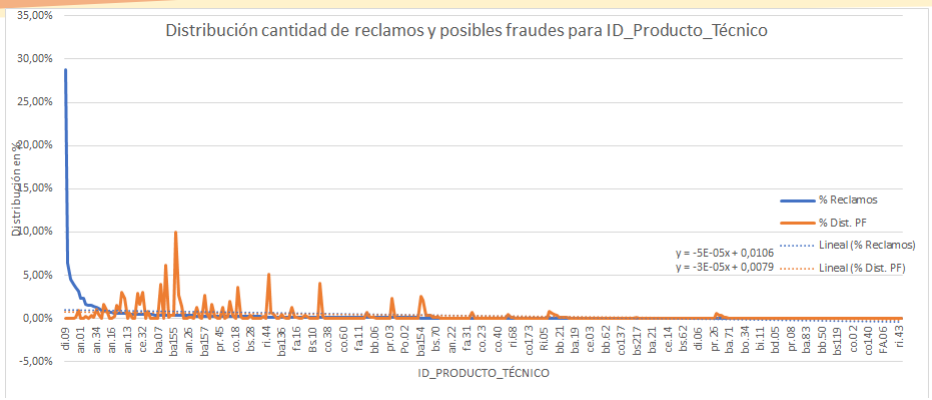


Gráfico 9. Distribución de reclamos y posibles fraudes para la variable producto técnico relacionadas para base reclamos y siniestros.

Al elaborar el coeficiente de correlación entre ambas variables del gráfico, dio un 0.009 lo que es una correlación ínfima casi nula.

Relacionando en base a BANCO, primero se identifica que hay 37 bancos distintos en la base Reclamos, mientras que en la base SiniestrosBD hay 38 BANCOS distintos, al cruzar los datos hay 9 bancos distintos en ambas bases.

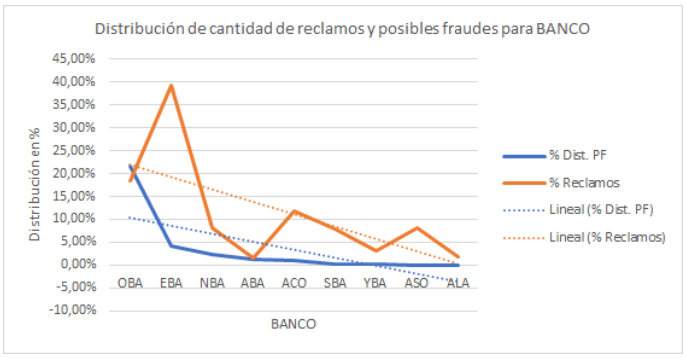


Gráfico 10. Distribución de posibles fraudes y reclamos para los bancos relacionados en base Siniestros y Reclamos.

Notar que los bancos que tienen más reclamos en la base Reclamos y que se encuentran en la base SiniestrosBD tienen una probabilidad más alta de ser POSIBLE FRAUDE, incluso tienen un coeficiente de correlación de 0.42. Es por ello que se concluye agregar nuevas variables binarizadas de estos bancos, que por cierto ya fueron consideradas en el análisis particular de la variable BANCO.

Adicionalmente, a través de DNI_ASEGURADO de la base Reclamos, se rescata la edad de la base Personas, y de este modo se determina si la edad de las personas que reclaman es una variable relacionada con POSIBLE FRAUDE. De esta misma forma, se repite el análisis con la variable Ciudad, de la base Personas. Ante ello solo se encuentra a una persona cuyo DNI_ASEGURADO está en ambas bases, lo cual no es representativo de la muestra de 9.723 instancias o filas de la base Reclamos ni tampoco de los 7.669 DNI_ASEGURADO distintos en la misma base.

En conclusión, de lo expuesto anteriormente, la relación más significativa de los reclamos y los posibles fraudes se encuentra en la variable BANCO, ya que entre más reclamos se halla de cada banco, más posible fraude contendrá este. No corre la misma interpretación para las demás variables disponibles de relacionar, mantienen una correlación cercana a cero, no pudiendo determinar si presentar más reclamos o menos influye en el hallazgo de más o menos posibilidad de fraude.

Con el fin de capturar esta interpretación de la variable BANCO, se realizaron variables binarias en el modelo predictivo que tomen el valor 1 para cuando la instancia tome los valores de dichos bancos.

ii) Hipótesis y supuestos sobre relación entre reclamos y devoluciones

Se tiende a interpretar que una persona cuyo reclamo ha sido rechazado en primera instancia debe reclamar más veces para ser aprobado, y que mientras no sea aprobado su reclamo, solicite por lo menos más devoluciones de la prima por el descontento que el rechazo del reclamo genera. Al analizar esta hipótesis, se logra demostrar que aquello ocurre sobre todo para los reclamos que terminaron siendo aprobados respecto a los aprobados con observación. Del siguiente gráfico se interpreta que aquellas personas que reclaman y logran ser aprobados como mínimo solicitan 4 devoluciones en promedio. Además, se interpreta que por cada vez que se reclame se pide en promedio una devolución de prima adicional aproximadamente.

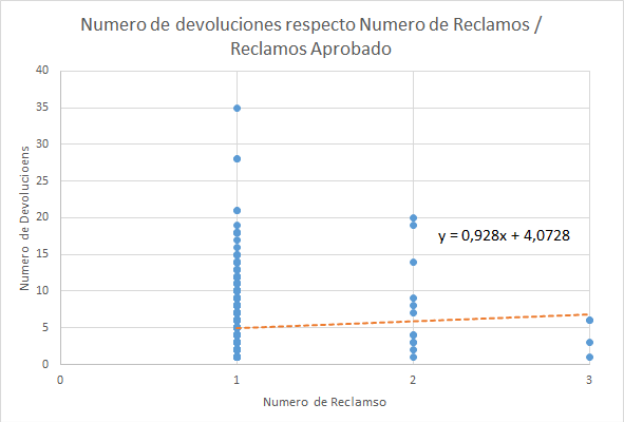


Gráfico 11. Numero de devoluciones respecto número de relamos/reclamos aprobado.

iii) Grado de relación entre devoluciones y fraude

Al relacionar las bases de siniestro y devoluciones a través de DNI_ASEGURADO, solo hay 2 instancias relacionadas y son clase no (no útil para determinar un grado de relación), por lo que se ahondará en relacionar las bases por la variable POLIZA_SEGURO.

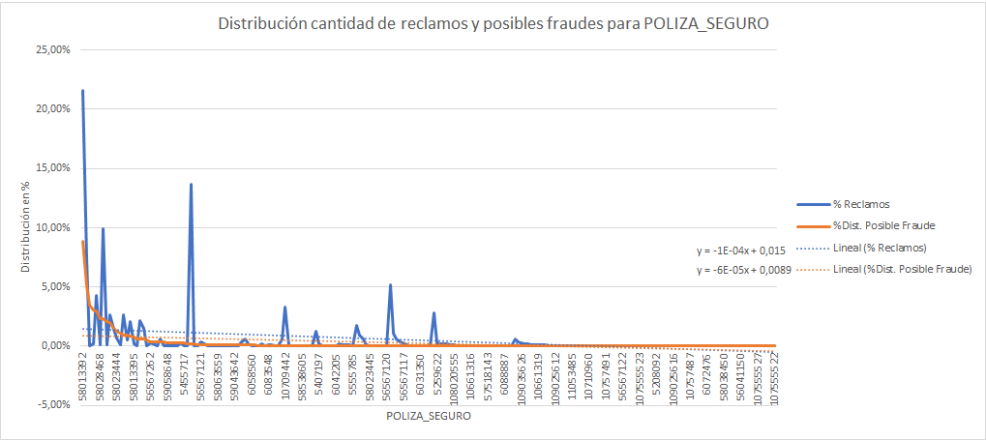


Gráfico 12. Distribución de reclamos y posibles fraudes para valores de POLIZA_SEGURO relacionadas para base devoluciones y siniestros.

Del gráfico x, se rescata que hay un comportamiento similar para la distribución de reclamos y de posibles fraudes para las distintas pólizas, esto es más notorio al considerar las líneas tendenciales ya que ambas presentan una pendiente similar solo que con coeficiente diferente. Al elaborar el coeficiente de correlación para estos, se halló un 0.71 de correlación, una relación bastante alta y que dicta que entre más devoluciones existan para una póliza, mayor fraude presenta este.

Para capturar este comportamiento, se creó una variable binaria que toma valor 1 cuando POLIZA_SEGURO toma uno de los valores de las 3 pólizas con más devoluciones, bajo la hipótesis que estos ayudarán al modelo a captar las instancias de posible fraude.

Solución Propuesta

Para hallar una solución respecto al problema de determinar qué situaciones detectadas como "posibles fraudes" deben ser analizados por la compañía aseguradora, se comienza por sentar las bases desde la detección de estos casos, particularmente, lograr predecirse para evitar pagarlas a los asegurados que las cometen. Se realiza en primera instancia un modelo predictivo que permite asignar qué solicitudes son consideradas fraudulentas. Con ese propósito, se comparan diferentes modelos, explicados en la sección "Modelo Predictivo", para determinar la configuración que entrega mejor precisión en esta detección, ya que de ello depende la optimización posterior.

Luego, se realiza un problema de optimización cuyo propósito es determinar cuáles de esas solicitudes detectadas como fraudulentas deben ser auditadas y revisadas para evitar su pago al cliente, principalmente porque los costos o desembolsos de pagar al cliente el monto que solicita son mayores a los de la contratación de abogados y otros costos asociados a revisar el caso.

Modelo Predictivo

Con el propósito de determinar qué modelo entrega una mejor precisión en cuanto a la detección, se procede investigar e identificar qué variación del modelo es el mejor de cada familia de métodos de Machine Learning. Con esto en mente, se comparan los desempeños de métodos Random Forest, Decision Tree, y Support Vector Machine. Para ello, y lograr optimizar cada uno se trabaja cambiando distintos parámetros de forma manual como porcentajes de balanceos, pero también ajustes de forma automática utilizando Grid Search utilizando muestras de 5.000, 10.000 y 15.000 datos para identificar con qué volumen de datos de entrenamiento cada modelo es más efectivo, dichas muestras son extraídas acorde a la distribución promedio de la variable 'Monto_Aprobado' de la base de predicción. Adicionalmente, las muestras se separan en 70% para entrenar el modelo, 15% de validación para ajustar cada método y encontrar los mejores candidatos de cada familia de métodos de machine learning y por último otro 15% de validación final o *Hold out data* para comparar y evaluar si efectivamente los modelos son precisos en la predicción y se ajustan para predecir nuevos datos o su desempeño en la primera validación solo se debió al sobreajuste.

En primer lugar, para la selección de muestras se extrajeron datos representativos de la base a predecir "Prediccion_v2" calculando la media de la variable monto aprobado de $554.700 y adhiriendo a esta un rango de desviación estándar de $100.000 para que el modelo aprenda de datos representativos, pero que al mismo tiempo no se sobreajuste tanto a los datos.

Posteriormente, se procede a la etapa de selección de variables. Para ello, en primer lugar, como ya se mostró, se realizó un análisis estadístico para seleccionar y generar variables que expliquen si una instancia es posible fraude o no, forma estadística pero también teórica. En este contexto, partiendo con una base de datos con 37 variables o columnas se llega a una base con 22, de las cuales se generan variables que mejor explicaban la clase a predecir generando una base con cerca de 50 variables predictivas. Aquello se realiza con el fin de acentuar el comportamiento de la clase y que el modelo predictivo aprenda de ello.

Con la última base, se construyen los modelos a ser optimizados. También se ejecuta la metodología de "Recursive Feature Elimination", basado en una eliminación "backward" y en función de un método random forest. Tras realizar esta metodología múltiples veces y de distintas formas se busca encontrar un subconjunto de variables predictoras que generen el mejor accuracy. En este sentido, como se observa en el siguiente gráfico se alcanza un ratio máximo con 29 variables.

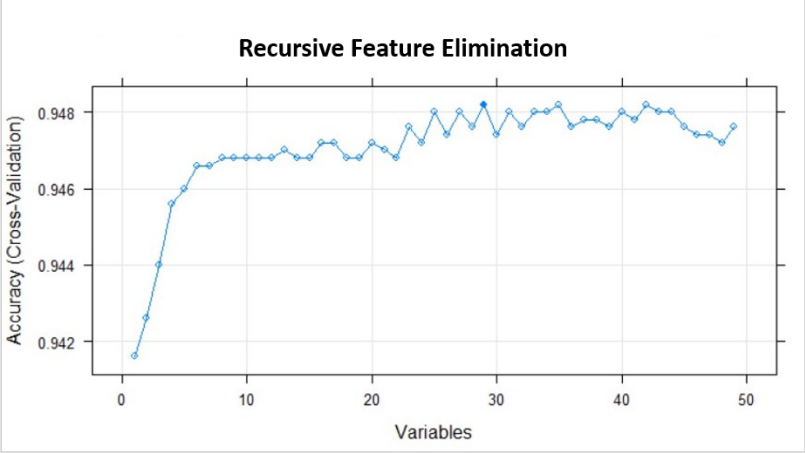


Gráfico 13. Recursive Feature elimination.

Ahora bien, para un problema con muestras desbalanceadas, solo considerar que el accuracy puede ser miope ya que con solo predecir los 0 (94% de la base de datos) se logra un buen accuracy, por ende también se toma como métrica importante y a mejorar la de sensitividad, sin sacrificar en grandes cantidades la especificidad. En ese sentido, el *trade off* entre estas métricas es, para un incremento de 80% a 90% de especificidad, la predicción correcta de 470 instancias (para la base de datos de 5000), en cambio ese mismo incremento para sensitividad se refleja en solo 30 instancias. Por ello, se ha decidido que la especificidad mínima a exigir a los modelos es de 90%.

De forma paralela al RFE, con el fin de capturar las variables más relevantes para construir el modelo, de la base con cerca de 50 variables se binarizan todos los valores de clase factor para obtener una base con solo datos de clase numérica. Luego, se balancea la base que contiene valores de CLASE, quedando en 50% con CLASE 'Posible Fraude' a través de balanceo ROSE, de esta manera su puede ver el comportamiento de las variables respecto a esta variable, con el propósito de discernir qué variables aportan a NO y Posible Fraude. Se realiza un análisis de correlación univariada de las variables predictivas respecto a la clase, y filtrando solo aquellas variables que tienen una correlación mayor al 7%. Este subconjunto de variables también es utilizado para la construcción de modelos predictivos. De esto, se obtienen 22 variables con correlación mayor al valor absoluto de ± 7%, para luego eliminar las que estén correlacionadas entre sí, con el fin de evitar los problemas de colinealidad (variables correlacionadas mayor a un 75%).

Ya con las variables seleccionadas después de este filtro por RFE y correlaciones, se procede a seleccionar dichas variables de la base nativa para poder balancear formalmente y que puedan entrar a los modelos. Para este balance se escogió un 30% de CLASE Posible Fraude, ya que así el modelo puede aprender a reconocer esta clase que podría ser difícil dada la naturaleza del 6% de esta para la base nativa y sin sacrificar en gran medida las métricas de especificidad, como se muestra en el gráfico siguiente.

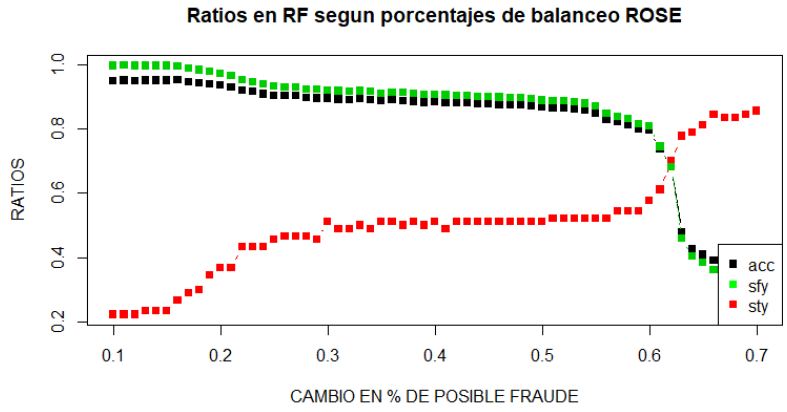


Gráfico 14. Balance de 10% a 70% de Posible Fraude entrenados y testeados en modelo Random Forest.

En síntesis, el procedimiento seguido para el modelo con el cual se trabajará posteriormente, es ilustrado en el siguiente esquema:

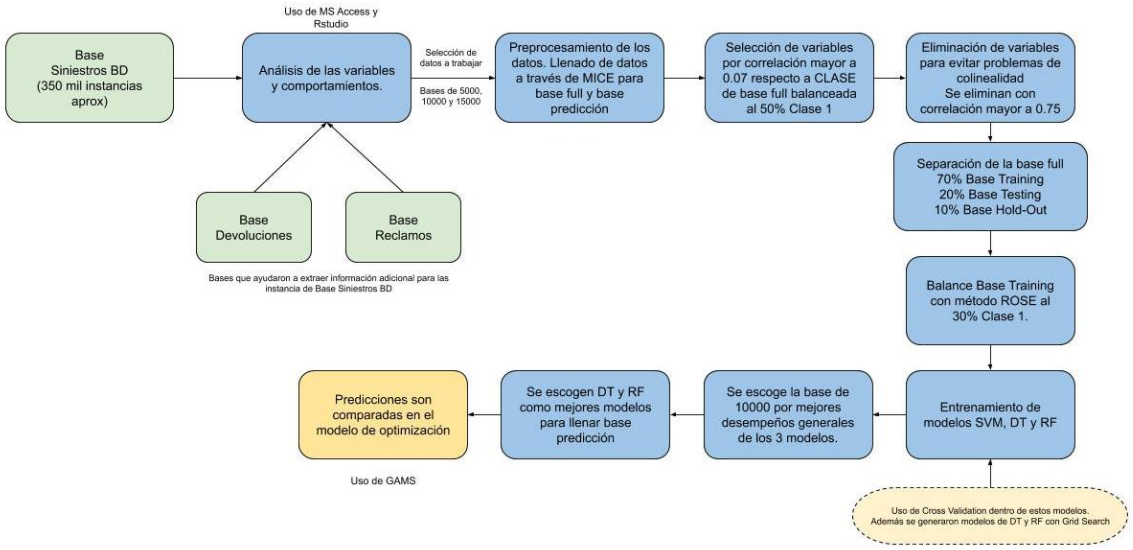


Ilustración 1. Síntesis procedimiento seguido para mejor modelo. Elaboración propia.

Tras haber entrenado y ajustado los modelos y observando su desempeño en la primera base de validación, se construye la curva ROC para cada familia de métodos de machine learning, como se observa el gráfico siguiente. De cada curva se selecciona el o los mejores modelos en función de que tan parecido haya sido su desempeño.

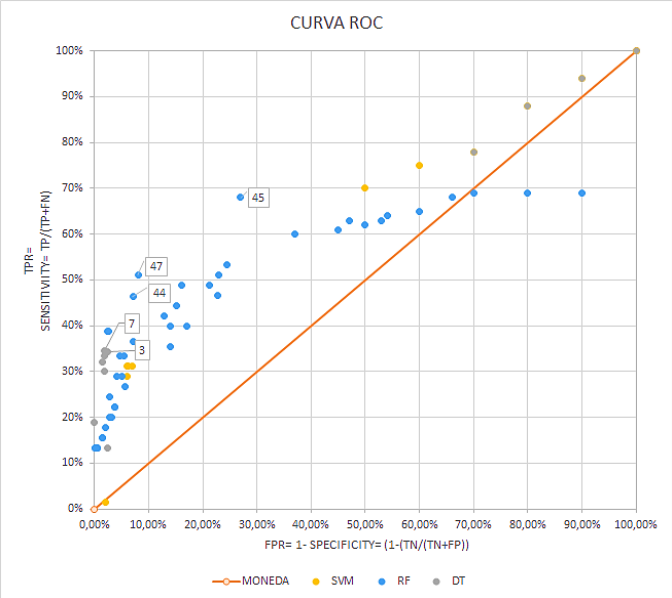


Gráfico 15. Curva ROC

Como antes se menciona, dado que en este proyecto tiene gran valor predecir correctamente instancias de clase "posible fraude" nuestro interés será encontrar un modelo cuyas predicciones tengan un alto Sensitivity. No obstante, no hay que olvidar que el actual problema presenta una naturaleza de clase desbalanceada con más instancias de clase "no" que "posible fraude", por lo que elegir un modelo con un buen Sensitivity a costa de un mal Specificity implicaría grandes magnitudes de errores. Aquello por ejemplo ocurriría al elegir un modelo muy riesgos como el número 45 que se muestra en el gráfico de curva ROC.

Como se observa en la curva, se selecciona el modelo 3 y 7 de la familia Decision Tree (DT), el modelo 44 y 47 de la familia Random Forest (RF), se descarta el modelo 45 por ser muy riesgoso ya que tiene un balanceo de 59%, por último se descarta la utilización de Support Vector Machine (SVM) por su mal desempeño. Cabe aclarar, que dado que Random Forest fue el modelo con mejor desempeño es que se visualizan muchos puntos azules. Para mayor detalle, se describen las características utilizadas para la construcción de los mejores modelos en la tabla posterior.

Modelos con mejor desempeño en validación				
Familia	**DT**		**Random Forest**	
N° punto en curva ROC	3	7	44	47
Grid Search	No	Si	No	No
Parámetros	default	default	ntrees=440 mtry=4	ntrees=440 mtry=4
Volumen Base	10.000	10.000	10.000	10.000
Balanceo	Rose 22%	Rose 22%	Rose 25%	Rose 30%
Variables	Todas	Sugeridas por Recursive Feature Elimination	Corte por correlación superior a 7% respecto a CLASE	Corte por correlación superior a 7% respecto a CLASE
Accuracy	93,87%	94,40%	92,00%	89,40%
Sensitivity	34,44%	34,44%	46,33%	51,11%
Specificity	97,60%	98,22%	92,83%	91,84%

Tabla 7. Modelos con mejor desempeño en validación.

Dado que estos modelos pueden tener un sobre ajuste, propio de la optimización del aprendizaje del modelo para predecir instancias de la validación, se procede a predecir una base Hold Out con 1.500 datos. Con ello, se busca identificar si los ratios de desempeño anteriormente señalados en la tabla son realmente representativos del desempeño de los modelos para cualquier instancia o solo para las contenidas en la validación. Los resultados de la predicción del Hold Out se presentan a continuación.

Modelos con mejor desempeño en Hold Out				
Familia	DT		Random Forest	
N° punto en curva ROC	3	7	44	47
Grid Search	No	Si	No	No
Parámetros	default	default	ntrees=440 mtry=4	ntrees=440 mtry=4
Volumen Base	10.000	10.000	10.000	10.000
Balanceo	Rose 22%	Rose 22%	Rose 25%	Rose 30%
Variables	Todas	Sugeridas por Recursive Feature Elimination	Corte por correlación superior a 7% respecto a CLASE	Corte por correlación superior a 7% respecto a CLASE
Accuracy	92,20%	92,80%	88,33%	87,67%
Sensitivity	21,11%	22,22%	35,56%	40,00%
Specificity	96,73%	97,30%	91,70%	90,71%

Tabla 8. Modelos con mejor desempeño en hold out data.

Como se puede ver de la tabla anterior, todos los ratios de todos los modelos disminuyen, concluyendo que existe sobre ajuste, pero no siempre en la misma magnitud. Independiente de aquello sobresale el modelo Random Forest numero 47 que no solo sigue presentando el mejor ratio de Sensitivity sino que también mantiene un ratio Specificity muy elevado sobre 90%. Si bien aquel modelo puede tener el Accuracy más bajo respecto a los otros modelos, su magnitud no es tan grande.

En base a lo descrito y bajo las actuales directrices de selección de modelos de Machine Learning se concluye seleccionar el modelo Random Forest, número 47 en la curva ROC.

En la sección *Anexos* se encuentra una tabla detalle del resultado obtenido por cada modelo que se aplicó para llegar al seleccionado.

Problema de Programación Lineal Optimización

La empresa aseguradora ya habiendo determinado qué casos son posibles fraudes con el modelo predictivo antes descrito, debe decidir si verifica que el caso es un fraude o no. Estudiar cada caso posee ciertos beneficios, ya que **se ahorraría pagar dinero a clientes por descubrir la falsedad de su solicitud**, junto a eliminar de su cartera de clientes a una persona que incurre en este tipo de prácticas. Sin embargo, para ello se incurren en costos, como la contratación de abogados, por lo que no siempre le es conveniente verificar todos los casos.

El problema de optimización pretende asignar qué casos deben ser efectivamente revisados, en concordancia con la combinación de casos que le genera una mayor utilidad a esta firma. Además, determinar la cantidad de abogados necesarios para el proceso. Matemáticamente, el problema se plantea como se muestra a continuación.

Previo: Supuestos Base

1. *Estar en la base Siniestros_BD implica estar cobrando el seguro. Cuando esta solicitud es rechaza, el asegurado puede reclamar.*
2. *Si se decide analizar un caso, se asume que el monto solicitado no será pagado a priori. Ergo, si se decide no analizar un caso se debe pagar el monto solicitado por el cobro aun cuando fuera un posible fraude.*
3. *Los abogados siempre cobran el mismo valor por hora y no se contratan abogados externos adicionales a los disponibles en la compañía (6). Solo se paga por las horas que trabajen analizando casos.*

I. Conjuntos

Para la resolución del problema se considerarán los conjuntos siguientes:

$b \in B = Conjunto\ de\ Bancos$
$i \in I = Conjunto\ tipos\ de\ seguro$
$t \in T = Conjunto\ dias$
$k \in K = Conjunto\ instancias\ (solicitud\ de\ cobro\ de\ seguros)$
$j \in J = Conjunto\ asegurados$

II. Nuevas Variables

La nueva variable (A), será clave para realizar la asignación que se desea, pues es la variable binaria que indica si efectivamente una instancia es investigada en un día particular.

$$A_{kt} = \begin{cases} 1\ si\ se\ investiga\ la\ instancia\ k\ en\ el\ dia\ t \\ 0\ si\ no\ se\ investiga\ la\ instancia\ k\ en\ el\ día\ t \end{cases}$$

Otra variable complementaria es G, que nos permitirá conocer la cantidad óptima de abogados que se enviarán a trabajar en los casos.

$$G_t = Cantidad\ de\ abogados\ a\ trabajar\ en\ el\ día\ t.$$

Una última variable complementaria agregada es I_{it} que representa la cantidad de seguros tipo i que se revisarán en el día t.

$$X_{it} = Cantidad\ de\ seguros\ tipo\ i\ que\ se\ analizarán\ en\ el\ día\ t.$$

III. Costos y Beneficios

Función Objetivo:

$$\boldsymbol{f} = \sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{t=1}^{T}(1-Y_k)\,M_k(1-P_{it})F_{ki}(1-A_{kt}) + \sum_{t=1}^{T}\sum_{k=1}^{K}M_kY_k\,A_{kt} + \sum_{j=1}^{j}\sum_{k=1}^{k}\sum_{b=1}^{b}N\,(1-L_b)V_{kb}O_{kj}A_{kt}$$

$$-\sum_{t=1}^{t}\sum_{k=1}^{k}\sum_{i=1}^{i}Q_iC\ F_{ki}A_{kt} - \sum_{t=1}^{t}\sum_{k=1}^{k}\sum_{i=1}^{i}M_kF_{ki}P_{it}A_{kt}\,(1-Y_k) - \sum_{j=1}^{j}\sum_{k=1}^{k}\sum_{b=1}^{b}N\,L_bV_{kb}O_{kj}\,A_{kt}$$

1) **Ahorro casos no fraude / Ahorro Teórico**

Corresponde al monto que se ahorra la compañía por aquellos seguros que nunca fueron reclamados por los clientes. M corresponde al monto, el cual es multiplicado por la cantidad de instancias no fraudes (Complemento del parámetro Y) y la probabilidades individuales de que estas no sean reclamadas (Complemento de P), además, se utiliza el parámetro F para relacionar la instancia a un tipo de seguro.

$$\sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{t=1}^{T}(1-Y_k)\,M_k(1-P_{it})F_{ki}\,(1-A_{kt})$$

2) **Ahorro casos fraude (Lo que habría que tenido que pagar)**

Es el monto que la empresa hubiese tenido que pagar a los asegurados por su solicitud de cobro, cuando estas fueron detectadas como fraudes. Este ahorro sucede cuando la empresa decide analizar el caso, bajo el supuesto 2.

$$\sum_{t=1}^{T}\sum_{k=1}^{K}M_k Y_k\,A_{kt}$$

3) **Devoluciones no solicitadas**

Cada cliente puede solicitar devolución de su dinero pagado a la compañía en cualquier momento. La ecuación entrega un monto promedio de dinero que se estima que los clientes no solicitarán de vuelta. Por ello, se multiplica la prima promedio por la probabilidad de que no se solicite devolución (Complemento de L), relacionando estos parámetros mediante O y V.

$$\sum_{j=1}^{j}\sum_{k=1}^{k}\sum_{b=1}^{b}N\,(1-L_b)V_{kb}O_{kj}A_{kt}$$

4) **Costos de investigación**

La investigación de las instancias está asociada a costos de abogados cuando se decide analizarlas. Esto corresponde a las horas (Q) de abogado que se requieren según tipo de caso por el valor hora de estos(C). La ecuación nos entregara un monto a incurrir.

$$\sum_{t=1}^{t}\sum_{k=1}^{k}\sum_{i=1}^{i}Q_i\;C\;A_{kt}F_{ki}$$

5) **Costo pago de reclamos**

Corresponde al desembolso por pagar las instancias que fueron reclamadas por clientes, independiente a si estas habían sido detectadas como fraudulentas o no.

$$\sum_{t=1}^{t}\sum_{k=1}^{k}\sum_{i=1}^{i} M_k F_{ki} P_{it} A_{kt} (1 - Y_k)$$

6) **Pago devoluciones**

Corresponde al monto pagado por las devoluciones de dinero que si fueron solicitadas. Al igual que en la anterior, se utiliza la prima promedio para determinar un monto aproximado de lo que será solicitado por los clientes.

$$\sum_{j=1}^{j}\sum_{k=1}^{k}\sum_{b=1}^{b} N\, L_b V_{kb} O_{kj}\, A_{kt}$$

IV. Restricciones

a) Cantidad horas Abogados

Existe un límite de horas disponibles para analizar, determinado por la cantidad de abogados a disposición. La ecuación entrega la cantidad de horas necesarias para la revisión de los casos fraudulentos, y restringe que estas deben ser menores a las horas abogado que se disponen, concretamente, 6 abogados por 9 horas al día[1].

$$\sum_{k=1}^{k}\sum_{i=1}^{i} Q_i F_{ki} A_{kt} \leq G_t * 9 \quad \forall t$$

b) Número de casos

Indica que la cantidad de casos a revisar para cada banco está sujeta a un límite superior y uno inferior, determinado arbitrariamente por la compañía. Es decir, los casos evaluados no pueden superar dichos límites por cada banco.

$$\sum_{k=1}^{k}\sum_{t=1}^{t} V_{kb=1} A_{kt} \leq 25 \quad \forall b = 1$$

$$\sum_{k=1}^{k}\sum_{t=1}^{t} V_{kb=10} A_{kt} \leq 30 \quad \forall b = 10$$

$$\sum_{k=1}^{k}\sum_{t=1}^{t} V_{kb=13} A_{kt} \leq 30 \quad \forall b = 13$$

$$\sum_{k=1}^{k}\sum_{t=1}^{t} V_{kb=4} A_{kt} \geq 30 \quad \forall b = 4$$

[1] *Aplicando Supuesto 3.*

$$\sum_{k=1}^{k}\sum_{t=1}^{t} V_{kb=18}A_{kt} \geq 20 \quad \forall b = 18$$

c) Tipos de Seguro

Corresponde a la cantidad máxima de instancias que pueden ser evaluadas respecto a cada tipo de seguro.

$$\sum_{t=1}^{t} X_{it} \leq R_i \ \ \forall i$$

d) 1 Revisión por instancia en un rango de 3 días

Se establece que 1 instancia particular solo puede ser evaluada 1 vez en un rango de 3 días, para lo cual se utiliza la variable A para indicar si una instancia fue revisada en un día específico.

$$\sum_{t=1}^{T} A_{kt} \leq 1 \quad \forall k$$

e) Máximo 6 abogados disponibles

La empresa solo dispone de 6 abogados para concretar las revisiones.

$$G_t \leq 6 \quad \forall t$$

f) Coherencia entre variables

$$\sum_{k=1}^{K} A_{kt}F_{ki} = X_{it} \quad \forall i, t$$

g) Naturaleza de variables

$$A_{kt} \in \{0,1\} \quad \forall k, t$$
$$G_t \in N\{0,6\} \quad \forall t$$
$$X_{it} \in N\{0,\infty\} \forall i, t$$

Análisis de Resultados

El resultado obtenido del problema antes explicado nos entrega la asignación de casos que la compañía debe llevar a análisis para obtener el máximo beneficio posible. En síntesis, esta asignación corresponde a lo siguiente:

	t=1	t=2	t=3
Casos a analizar	26	25	26
Abogados a contratar	6	6	6
Beneficio obtenido	$125.131.362		

Tabla 9. Asignación síntesis.

El primer factor de análisis es que, pese a que la cantidad de casos clase posible fraude era muy baja, el óptimo general se logra contratando a todos los abogados disponibles cada día.

Intuitivamente se interpreta que aun cuando un caso no pertenece al perfil de casos fraude, la mayoría de las veces conviene analizarlos debido a que pagarlos es un costo mayor que el de su revisión. Esta acción por parte de la compañía permitiría detectar la falsedad de una solicitud de cobro fuera de los parámetros establecidos por el modelo predictivo, sin embargo, constituye una decisión riesgosa, pues hay una alta probabilidad que la aseguradora deba gastar en análisis y además pagar el seguro, luego de confirmar su veracidad.

Según el resultado del modelo, la compañía deberá analizar 77 casos en el rango de días que se dedicará a la revisión, donde su capacidad de revisión está limitada por la cantidad de abogados disponibles cada día. Esta asignación otorga un beneficio total de $125.131.362 que es la diferencia entre los ahorros percibidos y los costos incurridos.

Dado que la interpretación de esta solución se vincula al *trade off* existente entre pagar una instancia y no hacerlo, a costa de un gasto por su análisis, se estima que la compañía aumentaría su beneficio (a priori) analizando todos los casos cuyo monto a pagar es mayor al de análisis, es decir, que de poder contratar más abogados probablemente lo haría. Sin embargo, el motivo de aquello es que no se está considerando lo que sucede luego de haber analizado los casos, donde se observaría que realmente habrán muchos casos que conllevarán un doble costo (pago y análisis) disminuyendo el beneficio final.

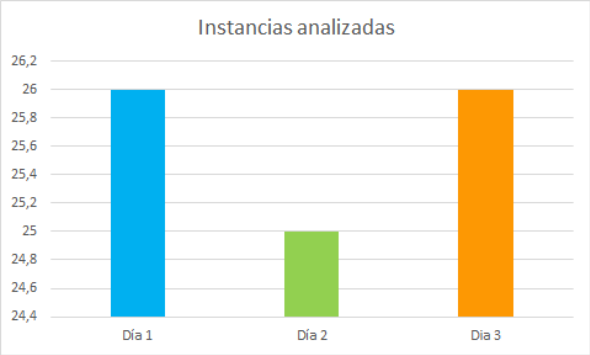


Gráfico 16. Instancias analizadas

Para los tipos de seguro, se observa que los casos llevados a análisis son mayoritariamente del tipo *despido (*i=5) y *enfermedad (i=1)*, lo cual puede deberse a que son este tipo de seguros los que conllevan un mayor pago al cliente, además de ser posiblemente los que poseen mayor frecuencia en las solicitudes. En síntesis, se analizan las cantidades según tipo de caso que se ilustran a continuación:

	t=1	t=2	t=3
i=1	2	4	2
i=2	2	3	0
i=5	22	18	24

Tabla 10. Asignación para cada dia t, según tipo de seguro.

Los seguros por despido son el grueso de la palestra de casos de análisis, donde a su vez corresponde al tipo de seguro que, en proporción, posee mayor cantidad de clase *posible fraude,* según el análisis exploratorio realizado en principio. Visualmente, se observa:

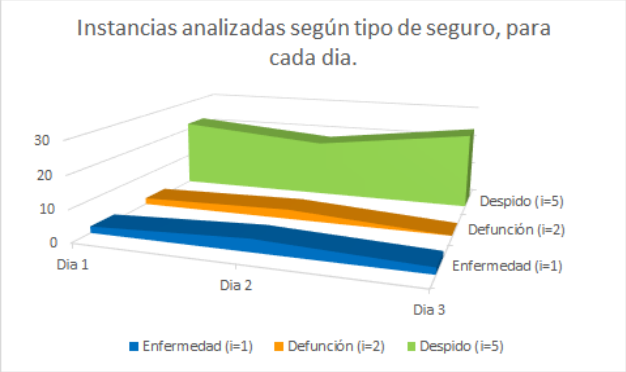


Gráfico 17. Instancias analizadas según tipo de seguro.

Sensibilización

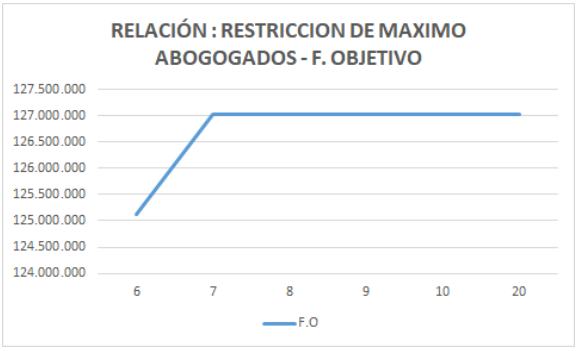
Para este análisis se consideran dos escenarios relevantes; el primero es observar cómo cambia el beneficio capturado en la función objetivo si se contratasen más abogados para ejecutar las revisiones. Mientras, el segundo es levantar las restricciones de mínimo y máximo asociadas a los bancos, que están establecidas por políticas de empresa.

En el caso de los abogados, se estima que, dado que no se cuenta con la información sobre lo determinado después del análisis, la cantidad de abogados a contratar será la máxima hasta que se logre cubrir la cantidad de casos cuyo monto a pagar sea mayor a los costos de análisis. Esto es porque para este problema, el análisis es siempre un ahorro de pago, sin embargo, cuando se tenga la información posterior al análisis, se podrá observar si dichos montos son realmente un ahorro.

Por otro lado, para las restricciones relativas a bancos por política de la empresa, se estima que el beneficio será mayor si estas no se consideran, ya que se eliminarían del análisis los casos que no requieren ser analizados, pero que se incluyeron para cumplir con la política.

a. Abogados a contratar: Se mueve la restricción relativa a abogados, donde se observa que siempre la empresa contrata el máximo. Para observar lo sucedido, se plantea la siguiente tabla:

Abogados	F.O
6	125.131.362,00
7	127.033.400,00
8	127.033.400,00
9	127.033.400,00
10	127.033.400,00
20	127.033.400,00



Se extrae que el óptimo de contratación de abogados, sin considerar el límite actual, es de 7 abogados, ya que, si bien la función objetivo se mantiene luego de aumentar la contratación, esto sucede porque no hay más casos para análisis. La empresa luego de 7 abogados contratados no incurre en más costos porque los abogados extra no están trabajando (capacidad ociosa), y la empresa paga a ellos por sus horas trabajadas, ergo, puede contratar infinitos abogados que no costarán dinero, dado que no estarán utilizando sus horas de trabajo.

La sensibilidad obtenida por este cambio es de $1.902.038.

b. **Restricción relativa a bancos:** La política de la empresa es tener un mínimo de 20 instancias para el banco RCO y 30 para el banco YRI. Esto fuerza al programa de optimización a asignar instancias para su revisión, cuando no necesitan ser revisadas, incurriendo en costos adicionales innecesarios. Primero, se reduce progresivamente la cantidad mínima de instancias para el banco RCO, donde se extrae que el beneficio óptimo aumenta al levantar dicha restricción, es decir, no revisando ningún caso de ese banco. Ello podría deberse a que las solicitudes de este banco no son lo suficientemente costosas como para que sea conveniente analizarlas por sobre pagarlas. Gráficamente esto es:

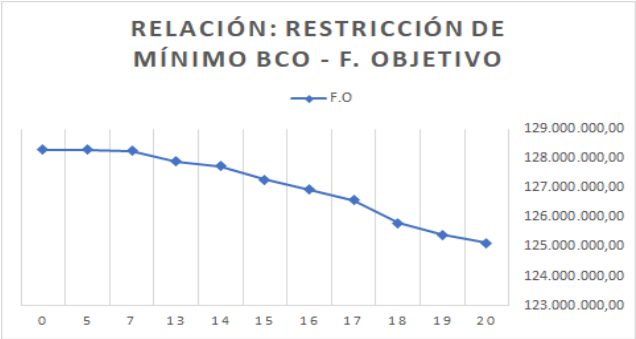


Gráfico 18. Levantamiento de restricción de mínimo banco RCO

Paralelamente, al reducir progresivamente la cantidad de instancias revisadas para el banco YRI, la función objetivo se mantiene constante, lo cual quiere decir que, aun levantando la restricción, el modelo analiza al menos los 30 casos que establece la política. Redundando en el análisis anterior, esto puede deberse a que las instancias de YRI son lo suficientemente costosas para que resulte conveniente analizarlas por sobre pagarlas, en la mayoría de los casos.

Para este análisis la sensibilidad es de $391.937 en promedio.

Preguntas para análisis PPL

i) ¿Cuántos casos con sospecha de fraude se están investigando y cuantos casos no se investigan pese a ser fraude? Las instancias llevadas a análisis de clase posible fraude son la 2, 68, 80, 82, 85, 102 y 137, que son en su totalidad las instancias detectadas como fraude por el modelo predictivo, es decir, se analizan todos los casos de clase 1 detectados.

ii) ¿En cuánto se estiman los beneficios y en cuanto el costo que considera todos las aristas del problema investigado? Para este punto, se separa la función objetivo en 2 variables (costos y ahorros) que se restan, así gams asignará el costo y ahorro que permiten maximizar el beneficio. Se observa como resultado:

Costo Óptimo = $46.842.580; Ingreso Óptimo = $171.973.942; Beneficio Obtenido= $125.131.362

iii) ¿Cuáles serían sus propuestas (2 al menos) respecto a sus resultados respecto al equipo de abogados y reclamos? Como fue analizado en la sección *Sensibilización* se sugiere a la empresa aumentar la cantidad de abogados disponibles a 7, ya que el beneficio obtenido es mayor. A su vez, respecto a los reclamos, se sugiere que se lleve a cabo un modelo de optimización particular para la base de reclamos, que permita analizar únicamente casos cuando sea realmente conveniente.

Conclusiones

En el estudio desarrollado en el presente documento se dio resolución a la problemática de una compañía de Seguros, cuyo dilema recae en la decisión de llevar a análisis las solicitudes de cobro de sus asegurados o no. Con este análisis, se resuelve de forma anexa la cantidad de abogados que la empresa requiere para concretar estas revisiones, y se desprende qué tipos de seguro son los que resultan mayormente conveniente de llevar a análisis.

Como primer paso, se elaboró un modelo de predicción basado en Random Forest, balanceado y utilizando variables extraídas de las bases de datos proporcionadas. Su determinación, permitió obtener la clase de los casos, siendo catalogados como posible fraude o no. Este proceso tenía por propósito llevar a estudio la hipótesis de que la revisión de casos que se clasifican como posible fraude, tienen un beneficio monetario significativo en la compañía. El procedimiento seguido para elaborar la predicción conllevó probar varios métodos de machine learning, balanceados en diferentes proporciones y bajo varias técnicas de mejora a estos, dando como resultado el descrito en la sección *Modelo Predictivo.*

A su vez, se planteó un problema de programación lineal (PPL) resuelto en el software de optimización GAMS para optimizar el beneficio monetario a través de la asignación de casos a analizar. Este planteamiento consideró las restricciones impuestas por la compañía y se trabajó para minimizar los costos mientras se maximiza el ahorro total.

Dado el resultado de aproximadamente 125 millones de pesos en beneficios, contratando a todos los abogados los 3 días dedicados a revisión, aun cuando la predicción determinó muy pocos casos de clase posible fraude, se confirma la hipótesis, es decir, efectivamente revisar los casos para verificar su falsedad conlleva un beneficio significativo para la empresa en la mayoría de los casos. Además, llevar a análisis casos que no fueron clasificados como tal, pero que cuyo pago es mayor al costo de analizarlos, también genera un ahorro importante para esta compañía.

Se desprende como conclusión que los siniestros de tipo despido resultan ser aquellos que conllevan mayor desembolso monetario, tanto por su necesidad de análisis como por el monto que implica su pago. Se recomendaría a la empresa reducir la cantidad de seguros de este tipo que se ofrecen a los clientes, ya que en ambos casos genera un desembolso importante de dinero.

Se cuestionan las restricciones asociadas a mínimo de casos por banco, ya que implican revisar casos que no tienen necesidad de ser revisados, incurriendo en costos innecesarios para la compañía. Mediante el análisis de sensibilización queda demostrado que levantando la restricción de mínimo para el banco RCO, la función objetivo logra capturar un beneficio mayor, así evitando incurrir en costos innecesarios, se logra mejorar el resultado obtenido.

Finalmente, se sostiene que el estudio realizado se encuentra sesgado al hecho de no conocer el resultado de la etapa posterior al análisis, es decir, al no saber si luego del análisis se verificó o no la falsedad de una solicitud, entonces no es posible determinar si el ahorro obtenido por este análisis fue materializado efectivamente para la compañía. Determinar, luego de haber incurrido en costos para analizar un caso, que este era verídico, cambia sustancialmente el resultado, pues implica que existirán varios casos donde ambas cosas sucederán, incurriendo en desembolsos mayores. Por dicho motivo, se deja como objeto de estudio futuro para la empresa, añadir el resultado posterior a la revisión, incorporando este factor al proceso de optimización ya desarrollado.

Anexos

Anexo 1. Análisis particular detallado

- **LOT_FECHA_RECEPCION:** Hace referencia a la fecha en que se recibe el siniestro en la compañía. Se evidencia que las instancias se concentran más en el año 2018 (54,23%) respecto al 2019 (45%), además hay solo una observación en 2005 de clase "NO". Al estudiar la distribución de las clases en los años 2018 y 2019 respecto al total de observaciones, se identifica que en el año 2018 hay más posibles fraudes, pero también, más clases "NO" respecto al 2019. Cabe mencionar, que se omite el año 2005 ya que posee solo 1 observación de clase "no".

Se calcula el porcentaje o probabilidad de posible fraude del total de observaciones del 2018 , 2019 y 2005 (%PF), en el siguiente gráfico 1. En él, se observa que el año 2018 tiene un ratio de un 8% (mayor al 6% de la base general), sumado a que concentra más posibles fraudes en volumen.

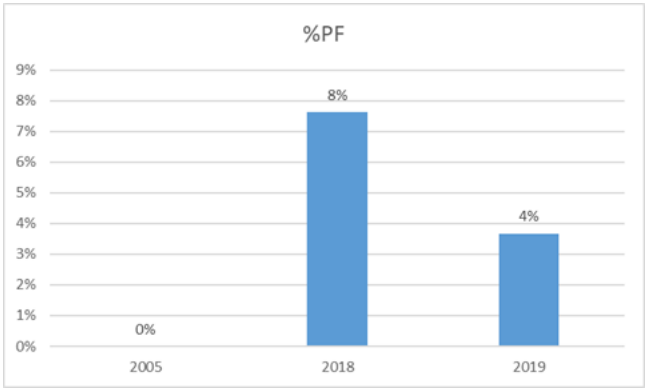


Gráfico 1. Porcentaje de POSIBLES FRAUDES para cada año de LOT_FECHA_RECEPCION.

Estudiando el porcentaje mensual de posible fraude (%PF) respecto al total de observaciones del determinado mes y año, como se ilustra en el posterior gráfico 2. En él, se identifica lo mencionado anteriormente respecto a los años, pero también una tendencia negativa, donde mientras más reciente sea la recepción del siniestro en la compañía, menos probable es que sea un posible fraude. No obstante, se identifica también que existen ciertos shocks en Septiembre y a principios de año.

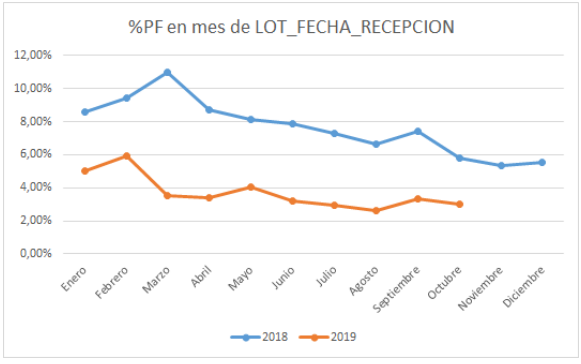


Gráfico 2. Porcentaje de POSIBLES FRAUDES para cada mes de LOT_FECHA_RECEPCION.

- **LOT_FECHA_RECEP_BENEFICIOS:** Analizando la variable, se descubre que tiene un comportamiento parecido a la variable anterior. En cuanto a su distribución en años y en distintas clases, se observa que 2018 tiene una mayor parte. Cabe mencionar, que también se identifica solo una observación de clase no para el año 2005. Respecto a su porcentaje de posible fraude (%PF) del total de observaciones para cada año, se concluye que las instancias del 2018 también son más probables de ser posibles fraudes que las del 2019. Mensualmente se identifica relativamente el mismo comportamiento que la variable LOT_FECHA_RECEPCION. De este modo se concluye que no vale la pena rescatar este comportamiento nuevamente.

- **POLIZA_SEGURO:** Para esta variable se encontraron 673 valores de pólizas, las cuales se crearon categorías con los dos primeros números de cada una como forma de clusterización, obteniendo 15 categorías diferentes y dando cuenta de 4 categorías que tenían un ratio de fraude mayor a la de la base de datos (6% aprox). Estas se consideran para ser binarizadas y ser incorporadas al modelo, sobre todo por las categorías 52 y 53 que son indicios de posibles fraudes.

CATEGORÍA POLIZA_SEGURO	NO	POSIBLE FRAUDE	% FRAUDE
10	38524	423	1,10%
11	79	4	4,80%
14	2	0	0,00%
50	420	18	4,10%
51	172	10	5,50%
52	2772	863	23,70%
53	324	228	41,30%
54	2824	256	8,30%
55	7995	319	3,80%
56	42273	2336	5,20%
57	24725	2676	9,80%
58	159291	11047	6,50%
59	31828	946	2,90%
60	3890	370	8,70%
61	457	1	0,20%

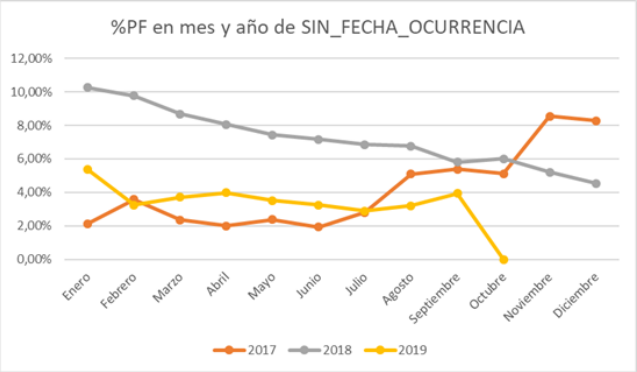
Tabla 2. Porcentaje de POSIBLES FRAUDES para cada año de LOT_FECHA_RECEPCION.

- **SIN_FECHA_OCURRENCIA:** Hace referencia a la fecha en que ocurre el siniestro, se identifica una distribución con mayor concentración de instancias en 2018 y 2019. Cabe indicar, que para siniestros ocurridos entre 2016 a 2003, junto al año 2001, 1990 y 1946 poseen un porcentaje de menos del 1% cada año, por tanto no se consideran para el análisis por su poca representatividad de la base de datos. Al observar las mismas cifras separando por clase y calculando el ratio o porcentaje de posible fraude del total de observaciones de cada año (%PF) se construye la siguiente tabla. Así, se interpreta que las instancias cuyos siniestros ocurren en el 2018 y 2017 son más probables a ser posibles fraudes.

AÑO_SFO	NO_OBS	PF_OBS	TOTAL	%PF
2017	19308	1494	20802	7,18%
2018	169200	13301	182501	7,29%

2019	122390	4664	127054	3,67%

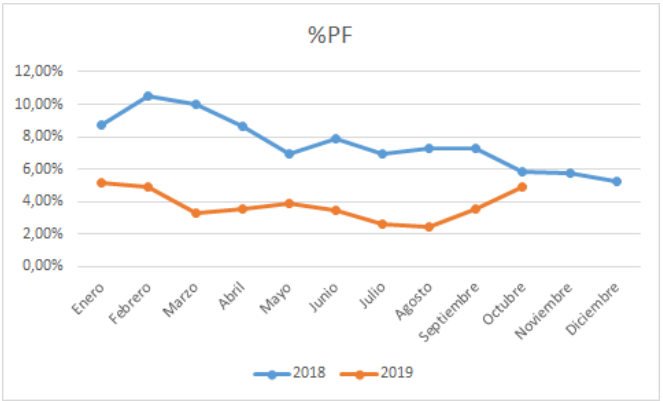
Analizando el comportamiento de posible fraude respecto al total de instancias del mes (%PF), se observa una tendencia alcista a fines del 2017 y bajista a lo largo del 2018 y 2019, como muestra el siguiente gráfico X. Asi instancias cuyos siniestros ocurren entre Noviembre de 2017 y Agosto de 2018 son más proclives a ser posible fraude.



- SIN_FEC_DENUNCIA: Esta se refiere a la fecha en que se denuncia el siniestro, su distribución se concentra ligeramente hacia 2018 que en 2019. Considerando la clase y el porcentaje de posible fraude para cada año (PF%) se construye la siguiente tabla. Así, se identifica que los siniestros denunciados en 2018 son más proclives a ser posibles fraude que en 2019.

AÑO_SFD	NO_OBS	PF_OBS	TOTAL	%PF
2018	171538	14100	185638	7,60%
2019	144039	5397	149436	3,61%

Mensualmente, las observaciones totales para la variable son relativamente estables, no evidenciado estacionalidad ni tendencia. En cambio, computando el porcentaje de posible fraude para cada mes y año (PF%), como muestra el siguiente gráfico X, se identifica una tendencia negativa de posible fraude mientras más reciente es la denuncia.Si bien este comportameinto ya es considerado por otras variables temporales, por su importancia teórica se mantendrá aunque estando sujeta a evaluaciones posteriores. Cabe mencionar, al igual que la variable anterior que los dos últimos meses de 2019 poseen muy pocas observaciones.



- **PRODUCTO:** Para esta variable, se tienen 1.400 valores diferentes, por lo que se procedió a categorizar las líneas de productos en los primeros dos dígitos de cada uno como forma de clusterización, obteniendo un total de 100 líneas. Al ordenar las líneas de productos de forma descendente por el ratio de fraude, saltan 2 líneas de productos muy indicadoras de fraude, la línea de producto EX y 59, destacadas por el mismo ratio y la cantidad absoluta (mayores a 200 instancias). Dicha variable de PRODUCTO será considerada en el modelo por su capacidad de predecir posibles fraudes dada la clasificación que se hizo como líneas de producto.

Línea PRODUCTO	NO	POSIBLE FRAUDE	% FRAUDE
EX	0	281	100,00%
20	6	12	66,67%
98	13	13	50,00%
59	813	335	29,18%
10	108	36	25,00%
52	18	5	21,74%
BA	8	2	20,00%
99	21	5	19,23%
53	13	3	18,75%
71	273	62	18,51%

- **SIN_FEC_INIC_VIG_PRODUCTO:** Se construye a siguiente tabla en que se ve mayor concentración en el 2016, Además, se identifica que los años 2003, 2014, 2015 y 2017 poseen una probabilidad mayor al 6% de posibles fraudes. Pero también, cabe considerar el año 2016 por el volumen de posible fraude que contiene. Los demás años por tener muy pocas observaciones o un porcentaje de posible fraude muy bajo no se consideran. Se concluye considerar esta variable ya que aporta nueva información que no aportan otras variables temporales.

AÑO_SFIVP	NO_OBS	PF_OBS	TOTAL	%PF
VACÍAS	7	0	7	0,00%
2000	1	0	1	0,00%
2001	5	0	5	0,00%
2002	458	9	467	1,93%
2003	213	94	307	30,62%
2004	147	51	198	25,76%
2005	264	22	286	7,69%
2006	678	55	733	7,50%
2007	2522	887	3409	26,02%
2008	2133	162	2295	7,06%
2009	2228	287	2515	11,41%
2010	7757	321	8078	3,97%
2011	1559	94	1653	5,69%
2012	15458	732	16190	4,52%
2013	29219	856	30075	2,85%
2014	23959	2018	25977	7,77%
2015	15743	1154	16897	6,83%

2016	147221	9270	156491	5,92%
2017	24296	2293	26589	8,62%
2018	37985	1065	39050	2,73%
2019	3724	127	3851	3,30%

- **DNI_ASEGURADO:** Se realizó un conteo de los posibles fraudes por cada DNI_ASEGURADO presente en la base de datos de SiniestrosBD y dividió por el total de fraudes presentes en esta, con el fin de ver la distribución de probabilidad de posibles fraudes para esta variable, en donde se presentó un DNI muy fraudulento, con 281 posibles fraudes de 281 registros, siendo '10900762512' el valor con probabilidad de 0,014% y 100% de tasa fraude. En ese sentido, también se considera los que tengan probabilidad mayor a 0,0015%, tomando así a los valores '109K0135012' y '10900122212'.

Así, se binarizan estos tres valores a nuevas variables para el preprocesamiento del modelo.

- **SIN_CUOTAS_COBERTURA:** Es la cantidad de cuotas que cubre el seguro contratado. Se procede a sustituir los ceros por uno que hacen referencia a una cuota. Aun cuando hay determinado número de cuotas que presenta mayor cantidad de casos de fraude, no se observa una tendencia de que mayor o menor número de cuotas genere más o menos probabilidad de ser detectado como fraude. Por esa razón, se generará una variable binaria que permitirá capturar a las cuotas menores a 6, que concentran mayor número de clase posible fraude.

SIN_CUOTAS_COBERTURA	Posible fraude
0	3278
1	200
12	103
13	68
3	4937
4	6534
5	1631
6	1452
7	487
8	807

- **SIN_FEC_INIC_VIG_SINIESTRADO**: Fecha en que se contrata el seguro. En la siguiente tabla se aprecia que las observaciones se concentran en los años 2017 y 2018, seguido por el 2016. Además, cabe mencionar que existen contratos firmados desde 2013 a 1900 pero no representan un porcentaje superior al 4% por si solos. De la tabla también se observa que los siniestros contratados el 2015 y 2016 tienen un mayor porcentaje de posible fraude, seguido por el año 2017 que a su vez es el año que posee mayor volumen de observaciones de posible fraude.

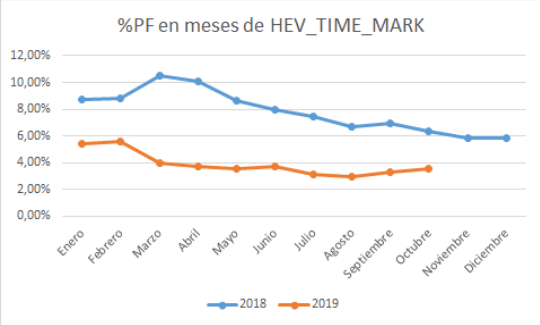
AÑO_SFIVS	NO_OBS	PF_OBS	TOTAL	%PF
2014	15483	1030	16513	6,24%
2015	27474	2090	29564	7,07%
2016	45524	3703	49227	7,52%
2017	82523	5630	88153	6,39%
2018	77917	3303	81220	4,07%

2019	15212	447	15659	2,85%

- **HEV_TIME_MARK:** Referente a la fecha en que se gestiona el siniestro en la compañía de seguros; se ve de la tabla que las instancias se concentran en los años 2018 y 2019. Se calculó el porcentaje de que una instancia sea posible fraude para el total de observaciones en cada año (%PF) en la que destaca el año 2018 con un gran porcentaje de posibles fraudes con un 8%. Adicionalmente, si bien el valor NULL posee un 9%, su volumen de observaciones es muy pequeño y además estas filas poseen NULL's en 10 variables más y la mayoría son de clase "NO", por lo que se eliminan por no aportar información relevante.

AÑO_HTM	OBS_NO	OBS_PF	TOTAL	%PF
2005	1	0	1	0%
2018	155986	13155	169141	8%
2019	159570	6340	165910	4%
NULL	20	2	22	9%

Mensualmente, se construye con la métrica PF del total de observaciones para un mes y año en particular. Cabe señalar, que las observaciones se distribuyen uniformemente en cada mes de cada año. En el gráfico, se una tendencia alcista desde enero hasta marzo de 2018 y luego tendencia negativa desde marzo 2018 hasta octubre del 2019 al igual que las fechas anteriores. Con ello, se infiere que mientras más reciente sea gestionado un siniestro menos probable es que se trate de un posible fraude. Dado que nuevamente se obtiene la misma tendencia negativa estará sujeta a eliminación en función de su aporte en la precisión del modelo.



- **HEV_DETAILS:** Al estudiar la variable identificamos que existen siete valores distintos incluyendo valores NULL. Además, las instancias se concentran mayormente en Dictamen Aprobado y Dictamen Rechazado. Al aperturar por clase , se calcula un total de observaciones y el porcentaje de posible fraude (%PF) para cada valor respecto al total de sus observaciones. A partir de ello, concluimos que es relevante considerar el valor Dictamen Aprobado ya que es el segundo porcentaje más alto de posible fraude, pero también, porque concentra un gran volumen de posibles fraudes. El resto de los valores se descarta por un volumen de observaciones y/o por un porcentaje de posible fraude bajo. Se concluye construir una nueva variable binaria sobre Dictamen Aprobado.

HEV_DETAILS	OBS_NO	OBS_PF	TOTAL	%PF
Anulado	35	5	40	12,50%
Dictamen Aprobado	214575	14152	228727	6,20%

Dictamen Pendiente	6865	341	7206	4,70%
Dictamen PreAprobado	159	3	162	1,90%
Dictamen Rechazado	93870	4992	98862	5,00%
Sin Dictamen	53	2	55	4%
NULL	20	2	22	9%

- SIN_ESTADO_ACTUAL: Referente al estado actual del siniestro después de la gestión de la compañía. Si bien se observan 11 valores distintos, estos son enclaustrados en 8, por la existencia de errores en la digitación. En cuanto a la distribución, esta se concentra en Pagados, En Evaluación y en Rechazado. Por otro lado, se construye la siguiente tabla donde se identifica que debemos considerar En Evaluación y Rechazado mediante variables binarias en nuestro modelo ya tienen porcentajes más altos (mayor o igual a 6%) y además representan una gran parte de las observaciones. Se descarta las demás por un bajo PF y/o TOTAL. Estos dos valores tienen sustento lógico si se considera que los siniestros que son posible fraude son más proclives a quedar rechazado o, por lo menos, en estado de evaluación.

SIN_ESTADO_ACTUAL	NO_OBS	PF_OBS	TOTAL	%PF
A Evaluar	16	1	17	5,90%
Aprobado	2866	63	2929	2,20%
Cerrado	15169	866	16035	5,40%
En Evaluacion	89336	6649	95985	6,90%
Pagado	156095	8726	164821	5,30%
Pendiente	5676	297	5973	5,00%
Rechazado	46418	2895	49313	5,90%
En Digitación	1	0	1	0,00%

- SIN_PAGO_ESPECIAL: Esta variable es igual a Y si se emite un pago con excepciones o N en caso contrario, la mayor parte de las instancias está asociada con el valor 'N' (99,57%) y menos de un 1% a Y. Al considerar ambas clases y el porcentaje de posible fraude, podemos ver que, si bien 'Y' no posee un gran volumen de datos, tiene un mayor porcentaje de posibles fraudes respecto a 'N' aunque no supera el 6%. Por ello, se concluye no considerar esta variable ya que no aporta mayor valor.

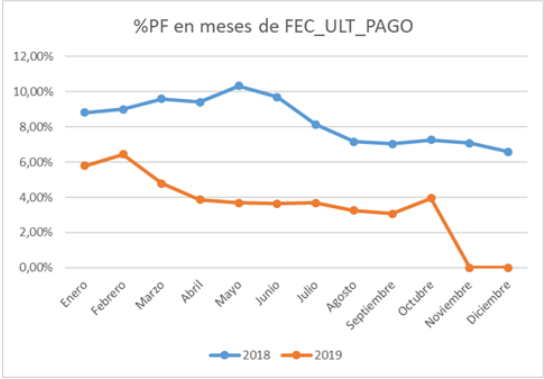
SIN_PAGO_ESPECIAL	NO_OBS	PF_OBS	TOTAL	%PF
N	314198	19436	333634	5,83%
Y	1379	61	1440	4,24%

- FEC_ULT_PAGO: Presenta solo 4 valores distintos, los cuales se distribuyen relativamente igual entre los valores NULL, año 2018 y año 2019. Se observa en la siguiente tabla X que a pesar de que el año 2019 tuvo mayores observaciones que el 2018, este último año posee una mayor concentración de posibles fraudes, doblando en porcentaje de posibles fraudes a 2019.

AÑO_FUP	NO_OBS	PF_OBS	TOTAL	%PF
2005	1	0	1	0,00%

2018	103990	9386	113376	8,28%
2019	112026	4841	116867	4,14%
NULL	99560	5270	104830	5,03%

Mensualmente, respecto a la misma métrica PF se puede observar en el siguiente gráfico una ligera tendencia positiva de enero a mayo del 2018 (sobre un 8%) y otra negativa desde mayo 2018 hasta diciembre 2019, ambos comportamientos que serán considerados para construir el modelo. Aquello se interpreta como, mientras más reciente haya sido la fecha de FEC_ULT_PAGO, menos probabilidad hay de que la instancia corresponda a un posible fraude. Cabe señalar que, si bien se observa un shock en octubre 2019, contiene muy pocos datos por lo que no se considera tan relevante.



- CUOTAS_PAG: Como esta variable se mide posterior a haber decidido el pago de un siniestro, no posee valor predictivo, motivo por el cual no se considerará en la predicción. Además teóricamente no existe una relación entre la cantidad de cuotas en que el siniestro es pagado y si este es fraude o no. Cabe añadir que cada cuota puede tener asociado distinto monto, por lo cual es más valioso capturar el monto que el número de cuotas pagado.

- MONTO_APROBADO: Para esta variable, que contiene gran cantidad de valores NULL, a través de ACCESS se cambió este valor a '0', ya que mayormente cuando aparecía este valor nulo es porque se rechaza el pago del seguro, o sea, se aprueba 0 pesos de pago, así se permitiría un análisis estadístico de esta variable respecto a la variable CLASE.
- LOT_EXTERNO: Esta variable hace referencia a si las gestiones fueron realizadas por una empresa externa (Y) o no (N), donde en primer lugar observamos que el 78% de la muestra corresponde a asuntos gestionados por empresas externas y el 22% restante de forma interna. Al aperturar por clase del total de casos gestionados por empresas externas, el 6% corresponde a la clase *posible fraude* mientras un 7% de los gestionados internamente corresponden a esta clase.

LOT_EXTERNO	NO FRAUDE	POSIBLE FRAUDE	TOTAL	%NO FRAUDE	%POSIBLE FRAUDE
N	69479	5072	74551	93%	7%
Y	246098	14425	260523	94%	6%

Ahora bien, analizando únicamente los casos que si son detectados como *posible fraude,* observamos que un 74% de estas detecciones son gestionados por empresas externas, lo cual puede verse explicado por la densidad de este factor en la muestra. A pesar de que puede tener relación

únicamente con que las gestiones de empresas externas son mayoritarias, es una realidad de la empresa que así sea, por lo que se considera relevante incluir la variable en el modelo.

LOT_EXTERNO	POSIBLE FRAUDE	%
N	5072	26%
Y	14425	74%
TOTAL	**19497**	**100%**

- **DNI_DENUNCIANTE:** Se realizó un análisis de los DNI_DENUNCIATE junto con la CLASE y da cuenta que un dni que tiene una tasa de fraude de 100% y con 281 instancias es el valor '10900762512', el cual es el mismo dni identificado en DNI_ASEGURADO en las mismas instancias. Por ello para no duplicar la variable no se considera binarizar. Por otro lado, si bien hay otros dnis que tiene tasa de fraude 100%, no son considerados por el número de instancias poco relevantes, menores a 50.

- **CAS_DESCRIPCION:** Se decide eliminar esta variable dado que captura la misma información que la variable anterior CAU_DESCRIPCIÓN.

- **FEC_PRIMERA_EVA:** En sí misma, la fecha en la que se realiza la primera evaluación no representa una variable de impacto en el modelo. Esto se debe a que esta fecha es dependiente de la decisión de la compañía aseguradora, basado en variables que no tienen relación con la situación a analizar, como tiempo y personal disponible para realizar la evaluación. Por este motivo se eliminará, sin perjuicio de que pueda utilizarse para derivar nuevas variables asociadas a rangos de tiempo.

- **TIPO_LIQUIDADOR:** Se refiere a si la gestión del siniestro fue realizada de manera interna o externa. Se extrae de las tablas de análisis que, si bien el impacto general de posibles fraudes en la muestra general es bajo respecto a esta variable, un 74% de los casos fraudulentos suceden en gestiones realizadas externamente, por lo que posee mayor probabilidad de ser un caso de fraude cuando se gestionan externamente los siniestros.

TIPO_LIQUIDADOR	NO FRAUDE	POSIBLE FRAUDE
EXTERNO	94,47%	5,53%
INTERNO	93,16%	6,84%

TIPO_LIQUIDADOR	POSIBLE FRAUDE	%PF
EXTERNO	14458	74,15%
INTERNO	5039	25,85%
TOTAL	**19497**	**100,00%**

La variable, sin embargo, captura la misma información que USER_LIQUIDADOR, en concreto, captura incluso menos información. Por ello, se elimina del modelo, manteniendo la mencionada como se indica en su análisis.

Anexo 2. Cuadro comparativo de resultados para modelos evaluados.

			BASE 5000			BASE 10000			BASE 15000		
			Validación_1	Validación_2	Testing	Validación_1	Validación_2	Testing	Validación_1	Validación_2	Testing
DT	Normal	ACC				95,13%			94,58%		
		SENS				19,00%			9,60%		
		SPEC				100,00%			100,00%		
	Balanceado	ACC				93,87%	92,20%		93,56%		
		SENS				34,44%	21,11%		20,00%		
		SPEC				97,60%	96,73%		98,25%		
	Aplicado	ACC				94,13%	92,80%		93,69%	93,60%	93,64%
		SENS				30,00%	21,11%		22,20%	23,70%	22,96%
		SPEC				98,20%	97,37%		98,25%	98,06%	98,15%
	Grid Search Balanceado	ACC	93,00%			94,27%	93,20%	93,73%			
		SENS	13,33%			33,33%	22,22%	27,70%			
		SPEC	97,59%			98,15%	97,70%	97,90%			
	Grid Search Aplicado	ACC				94,40%	92,80%	93,60%			
		SENS				34,44%	22,22%	28,33%			
		SPEC				98,22%	97,30%	97,77%			
	SENS -APLICADO	ACC				94,53%	92,28%	93,70%			
		SENS				32,20%	22,22%	27,20%			
		SPEC				98,50%	97,30%	97,94%			
SVM	Normal	ACC									
		SENS									
		SPEC									
	Balanceado	ACC									
		SENS									
		SPEC									
	Aplicado	ACC							92,70%	93,02%	92,80%
		SENS							1,48%	6,60%	4,07%
		SPEC							98,50%	98,50%	98,53%
	Grid Search Balanceado	ACC	89,33%								
		SENS	31,11%								
		SPEC	93,05%								
	Grid Search Aplicado	ACC									
		SENS									
		SPEC									
	SENS -APLICADO	ACC									
		SENS									
		SPEC									
RANDOM FOREST	Normal	ACC									
		SENS									
		SPEC									
	Balanceado	ACC				93,93%	92,20%	93,07%	92,71%	92,93%	
		SENS				38,88%	21,11%	28,80%	22,96%	25,92%	
		SPEC				97,44%	96,73%	97,16%	97,16%	97,21%	

	Aplicado	ACC		94,06%	92,87%	93,30%	92,76%	93,02%
		SENS		38,88%	26,66%	31,67%	22,96%	28,15%
		SPEC		97,58%	97,09%	97,23%	97,21%	97,16%
	Grid Search Balanceado	ACC						
		SENS						
		SPEC						
	Grid Search Aplicado	ACC						
		SENS						
		SPEC						
	SENS -APLICADO	ACC						
		SENS						
		SPEC						