**Detección de fraude en reclamación de seguros**

**Introducción**

El fraude se define como toda actividad ilegal mediante la cual una o varias personas intentan obtener beneficios económicos, para ellos mismos o para un tercero, mediante el engaño, dejando a su contraparte perjudicada. Esta conducta se da principalmente debido a las asimetrías que existen en la información, como por ejemplo en la veracidad de los sucesos y la dificultad para una de las partes la investigación y/o comprobación de los hechos. Saddler define el fraude como "todo acto intencional perpetrado por un ser humano, utilizando el engaño con el fin de obtener un beneficio personal" (Saddler, 2013).

La anterior definición se puede limitar al sector seguros, definiéndose entonces como todo acto que busque beneficiarse económicamente de un contrato de seguros mediante el engaño a las compañías aseguradoras. Dichos actos se dan cuando el asegurado, aprovechándose de la asimetría de la información y de la transferencia del riesgo a las compañías aseguradoras, planea/fabrica y ejecuta un siniestro o exagera las consecuencias del mismo, para que mediante esto, pueda obtener la indemnización que se pactó en la póliza de seguro.

Colombia se encuentra dentro de los países de Latinoamérica que más se ven afectado por este fenómeno, llegando a tener cifras alarmantes en donde 7 de cada 10 siniestros tienen algún indicio de fraude, pero en donde solamente se llega a comprobar este acto ilícito en el 7% de los casos sospechosos.

Existen diferentes estimaciones de este flagelo en la sociedad colombiana para los diferentes ramos, Fasecolda por su parte estima que para el ramo SOAT (Seguro Obligatorio de Accidentes de Tránsito) el costo de este delito alcanza más de 100.000 millones de pesos, por otro lado, la fiscalía puso en descubierto en abril de 2019 una red estafadora en el ramo Salud, que alcanzó a realizar estafas por 90.000 millones de pesos, el INIF (Instituto Nacional de Investigación y Prevención del Fraude) por su parte logró advertir a las aseguradoras en el 2017 que el 8% de los casos investigados fueron en efecto fraude ayudándoles a conseguir a las aseguradoras de este modo un ahorro de 11.614 millones al no tener que realizar pagos ilegítimos.

**Objetivo del proyecto**

Se espera que este proyecto sea de utilidad para ayudar a identificar las reclamaciones fraudulentas con anterioridad a los pagos de sus correspondientes siniestros y de este modo contribuir a la disminución de pérdidas para las compañías de seguros en reclamos ilegítimos, incrementando en últimas las utilidades de la compañía, del sector, de la economía y con suerte la disminución de las primas que cobran a sus asegurados, dado los menores gastos de operación.

Para lo anterior, se espera de este proyecto, poner a la disposición de las entidades aseguradoras una herramienta que les sirva para identificar reclamaciones que tengan altas probabilidades de haber sido fraudulentas para que de este modo las aseguradoras puedan enfocar sus recursos de investigación y denuncia de fraude en aquellas reclamaciones que a pesar de sus cuantías tienen altos indicios de fraude, para de este modo evitar que se siga normalizando este tipo de actos en el sector.

**Materiales y métodos**

Lastimosamente los datos para llevar a cabo este proyecto en el sector asegurador colombiano son de uso exclusivo de las compañías de seguros, Fasecolda e INIF. Esto porque, en caso de que se hiciera pública la información del fraude cometido en las diferentes aseguradoras; podría atraer de forma masiva estafadores llevándolas a atravesar una oleada de personas que intenten aprovecharse de este acto ilegal.

Dada la confidencialidad de los datos, es imposible realizar una investigación de este fenómeno en el ámbito colombiano. Por lo anterior se trabajó con base a información de la misma categoría / problemática que ha sido previamente difundida por una o varias aseguradoras estadounidense.

Por lo anterior para avanzar en el proyecto, se espera trabajar sobre este set de datos, y dejar la metodología de detección lista para su implementación cuando un conjunto de datos del sector colombiano esté disponible.

Teniendo en cuenta que el objetivo de este proyecto es la construcción de un modelo de machine learning que sea capaz de clasificar, y por añadidura encontrar a priori, los casos de fraude que se cometen en las reclamaciones de seguros; es necesario un dataset que contenga información acerca de los fraudes que se cometieron en una compañía de seguros, los cuales hayan sido documentados con sus respectivas etiquetas, comisión o no de hechos fraudulentos, junto a algunas características de sus clientes y del incidente, para con ello poder alimentar y entrenar el modelo de machine learning.

Como se verá más adelante, en el análisis de la variable fraud\_reported de este dataset, son más frecuentes los casos en los que no se comete fraude comparados contra los que sí, haciendo más difícil encontrar patrones que nos lleven a una cauterización de los casos fraudulentos más generalizada, dado el bajo número de entradas que tenemos con estas características. Este tipo de problemas y en general el problema de clasificación de fraude, se les conoce como problema de imbalance data, donde un tipo de datos ocurren más frecuentemente comparado con otros. La descripción y el entendimiento adecuado de los datos se hace crucial entonces, pues mediante este se podrá determinar qué variables deben incluirse o no dentro de la construcción del modelo, para que este a su vez haga un mejor trabajo en la clasificación.

**Contextualización de la data**

Para lograr cumplir con el objetivo planteado en este proyecto se tendrá que usar un dataset estadounidense proveído en un reto de Kaggle; el cual contiene 1000 registros de reclamaciones del seguro, sus características, así como una variable label donde nos informa del incurrimiento de fraude o no por parte del cliente. Las entradas de este dataset fueron tomadas desde el 01 de enero de 2015 hasta 01 de marzo del 2015, es decir, corresponde a 2 meses de información.

El dataset cuenta con 39 variables entre cualitativas y cuantitativas que describen las características que tiene una póliza, las del incidente que llevó a la reclamación y del cliente que la adquirió dicha póliza. Esta recolecta información del cliente como lo es meses de vinculación, edad, el género, la ocupación, los hobbies, la relación con el beneficiario; de las caracteristicas del incidente como año del auto, la existencia reporte policial, el daño a propiedad, cantidad de testigos presentes, numero de lesionados, entre otras; y las caracteristicas de la póliza información como número de la misma, fecha de vinculación, estado donde la adquirió, prima de la póliza entre otras.

**Análisis estadístico de los datos: Principales variables de interés**

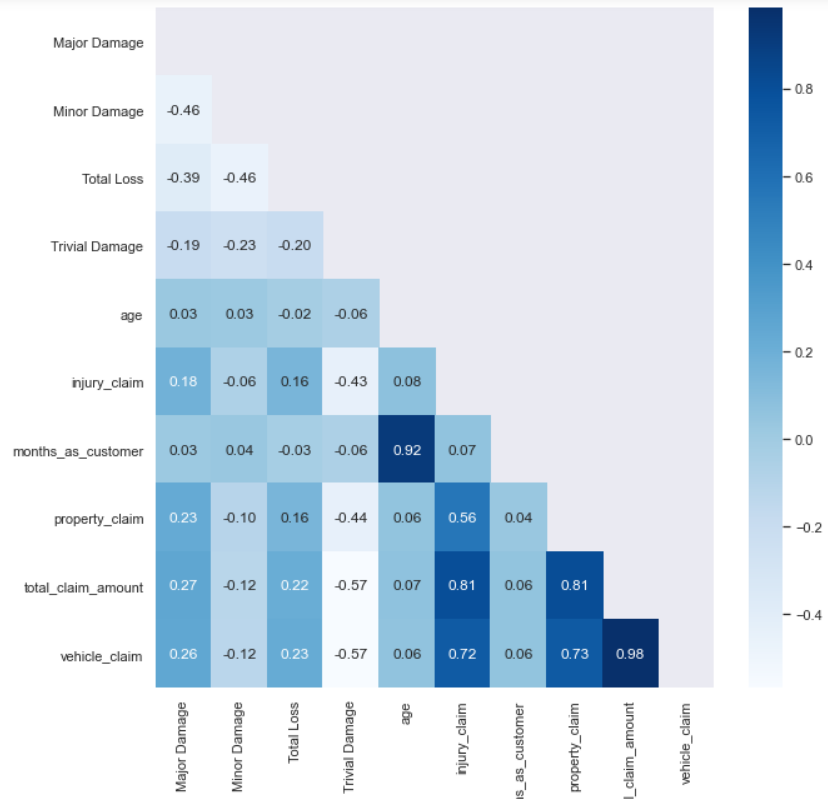
Mediante la variable fraud\_reported o la variable label, se observa el problema de desbalance de los datos, representando el número de casos que se han reportado como fraude. El gráfico de barras muestra que en la mayoría de los casos no existió fraude (N), con valores que van de 250 casos donde identificó como fraude( Y) y 750 sin reporte de fraude(N), es decir, de 25% vs 75% de los casos. Con ello se ve con más claridad el tipo de problemática que estamos abordando, en donde los casos de reclamaciones ilegítimas se ven representados en menor medida en nuestro dataset, dado que dificulta nuestro proceso de caracterización debido al bajo numero de muestras que tenemos del mismo.



Mediante el heatmap se evalua la correlacion entre 6 variales numéricas continuas y 1 variable categorica que fue dummificada para esta correlación. Por parte de las variables numéricas se analizó edad del asegurado, valor de reclamacion por lesiones, meses como clientes, valor de reclamacion por propiedad, valor de reclamación total y reclamacion del vehiculo; y por parte de la variable categorica se utilizó severidad del incidente.

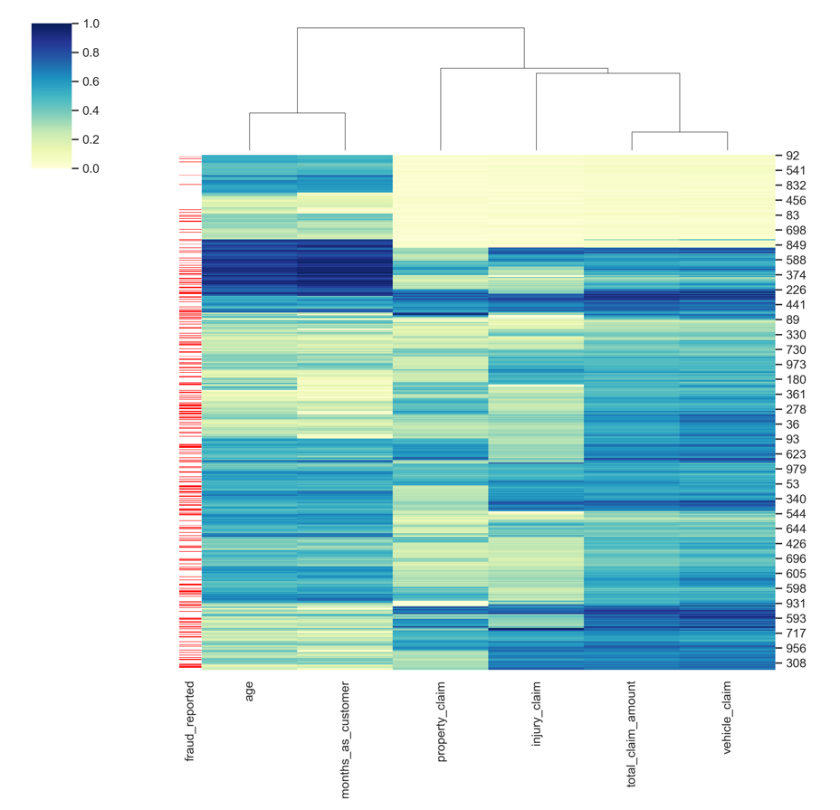
Se encontró que existe una alta correlación entre el valor total de la reclamación con el valor reclamado por el automóvil, por propiedad y por lesión con una correlación mayor a 0.81, lo que es de esperarse debido a que estas son unas subcategorías del valor total de reclamación. Estas variables a pesar de ser el desagregado de total del reclamo no se eliminaron del análisis porque aportan información individualmente, que no sería posible contrastar de manera agrupada mediante la variable total de reclamación. Del mismo modo, hay una correlación entre estas subcategorías de la reclamacion, que oscilan entre el 0.56 hasta 0.73.

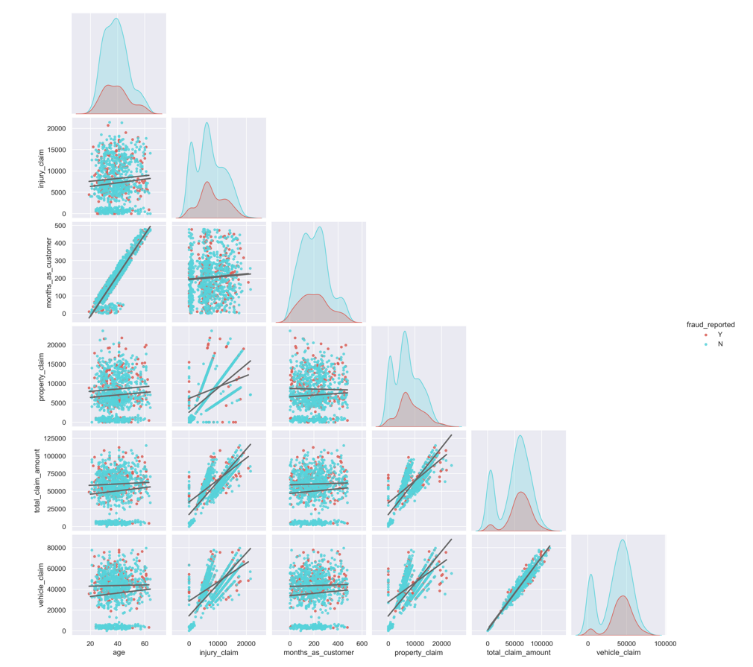
Otra importante correlacción que se observa, está entre la edad del asegurador y los meses que lleva como cliente, esto puede ser debido a que entre más edad tiene una persona, más se preocupa por el detrimento que pueda sufrir su patrimonio mediante eventos fortuitos.

finalmente, se observan ciertas correlaciones de los montos por reclamación con la variable dummy, que agrupa las categorias de daño mayor, daño menor, daño trivial y pérdida total.

El clustermap por su lado, nos ayuda a observar cuales son las variables se parecen y cuales difieren más. Para ello realiza una agrupación mediante el método de Hierarchical Aglomerative Clustering (HAC), el cual consiste ir generando clusters con las variables que menor distancia tengan. A su vez este método utiliza dendogramas que se observan en la parte superior del nuestra figura, los cuales muestran la manera en que fueron formados los grupos. Se ve que las variables que más relación tienen son reclamación por vehículo con total de reclamación, seguido por edad y meses como cliente. El método fue agrupando el resto de variables hasta dejar un solo cluster, en donde se encuentran todas las variables. Se observa que la agrupación se hizo por un lado con los valores de la reclamación, por vehiculo, lesión, propiedad y total, y por el otro lado una agrupación de edad y meses como clientes.

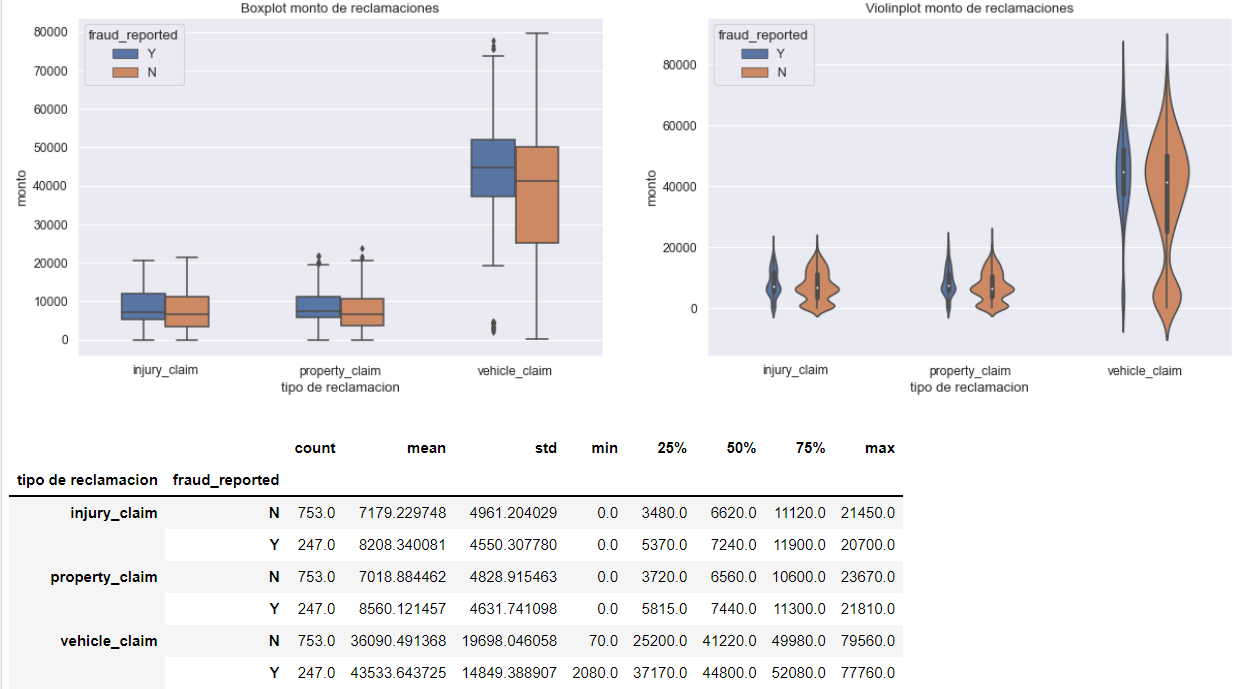
De igual manera se intentó colocar una bandera, que identifica a los casos en donde se cometió fraude, correspondiente a las lineas rojas que se encuentran a la izquierda, para poder identificar asi una relación apriori que tuvieran los defraudadores con las variables de edad, meses como cliente, y los montos de la reclamación. Sin embargo, se observa que la relación no es clara, y que los defraudadores no tienen un patrón bien definido, al menos con respecto a estas variables, pues las lineas de bandera representadas en rojo a la izquierda no se agrupan y por el contrario, se ven dispersas a través de todo el eje Y.



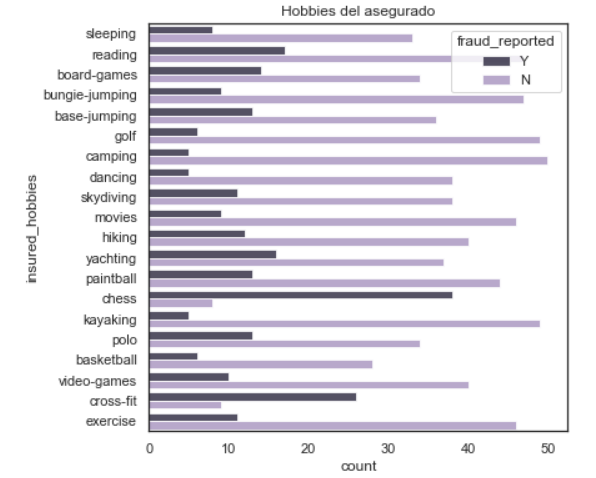
Finalmente, se dividieron las variables de edad, meses como cliente y montos de reclamación por la variable label, es decir, en casos en donde se reportó fraude y casos en los que no se reportó. Todo lo anterior con la finalidad de identificar si existía una correlación más evidente en los casos donde estuvo presente el fraude. Como se observa mediante los puntos marcados en rojo, la relación se hace mucho más difusa, pues los casos de fraude parecen tener un comportamiento aleatorio y poco consistente.

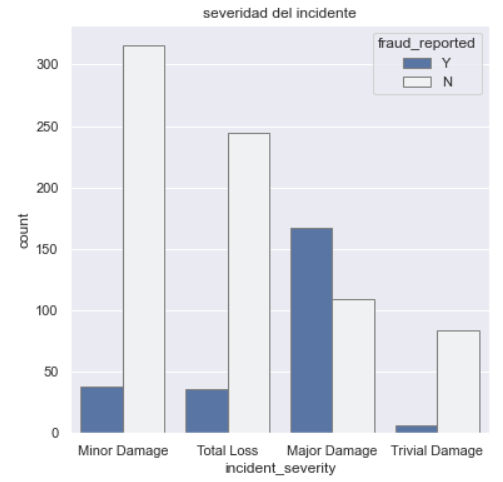
Al revisar el boxplot y violinplot discriminado por comisión y no comisión de fraude, podemos observar que existe una tendencia de las personas fraudulentas a hacer reclamaciones por valores superiores a los que reclamaría o reclamó una persona que no cometió fraude, esto se ve claramente en la los valores de los percentiles 25, 50 y 75, donde se ve que sus valores son sobrepasan los valores de reclamación que se hicieron por parte de las personas que no cometieron fraude.

Adicional a lo anterior, parece que los defraudadores tienen una tendencia a ubicar sus reclamaciones en montos centrales, pues se ve que la gran mayoría de los datos se ubican entre los USD 22000 y USD 100000, lo cual podría estar dado por un pensamiento de no levantar sospechas que conlleven a futuras investigaciones por hacer reclamaciones por valores excesivamente altos, ni su interés en hacer fraude en reclamaciones por valores excesivamente bajos.

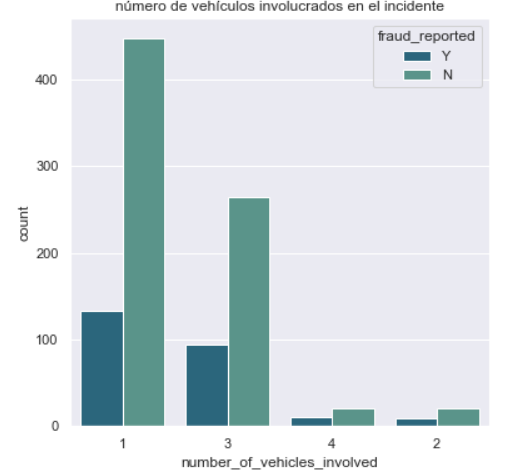


Los hobbies que reportó el asegurado tiene una alta incidencia o correlación con la comisión de fraude, especificamente en dos de sus valores. Se observa mediante esta gráfica de barras que los clientes que mostraron como hobbies el ajedrez y cross fit, tienen un alto porcentaje de comisión de fraude, inclusive llegando a ser las reclamaciones por parte de defraudadores superiores a los no defraudadores para estos dos hobbies. Se observa que el 80% de las personas que tienen como hobby el ajedrez, y que el 71% de los que practican cross-fit llegaron a cometer fraude. Por el resto de categorías, parece que se mantienen las proporciones de fraudulentos y no fraudulentos a través de los diferentes hobbies.

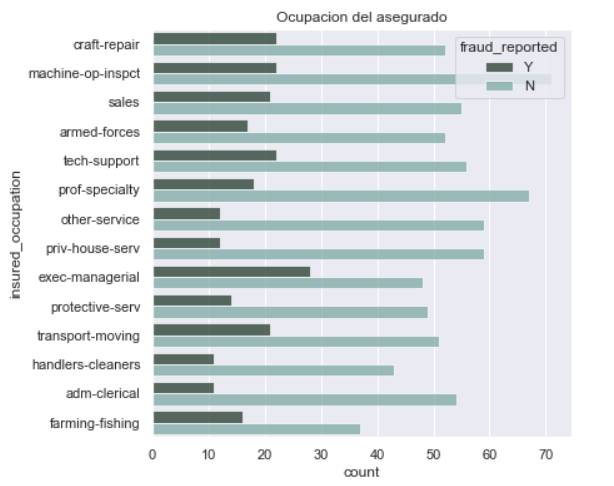


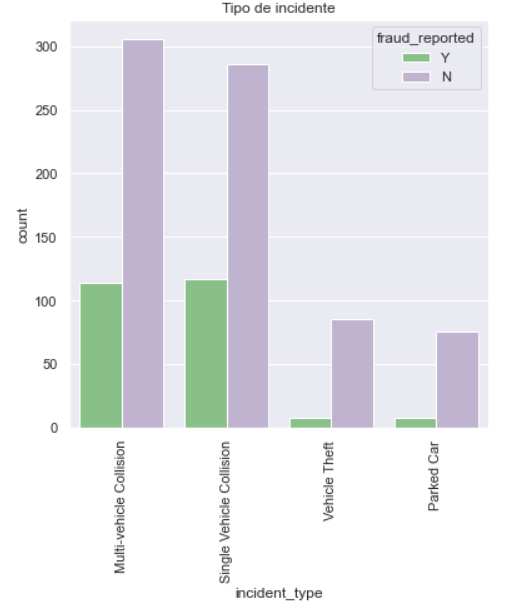
La variable severidad del incidente es una variable que podría llegar a explicar en gran medida la comisión de fraude, dado que tiene tiene categorias que correlacionan fuertemente con la comisión del fraude. Especificamente, se observa que los defraudadores reportan en esta variable la existencia de un daño mayor, al menos en este dataset, llegando a ser la cantidad de personas que cometieron fraude superiores a los que no con una proporción de 62% contra el 38%. Esta variable, al igual que monto de la reclamación, muestra que los defraudadores prefieren apuntar por los valores medios-altos, con el fin de no levantar sospecha que lleven a investigaciones en caso que lo hicieran por incidentes de severidad de grado de perdida total, pero tampoco tienen interés en reclamaciones de daños triviales o menores, lo cual se logra observar en las proporciones que tienen los defraudadores contra los casos legítimos en este tipo de daño donde no alcanza a superar el 15%.

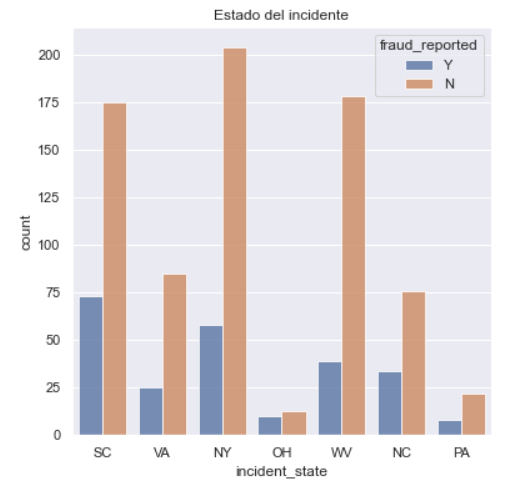
**Análisis estadístico de los datos: Otras variables de interés**

Mediante el gráfico de barras para la variable numero de vehiculos involucrados, se puede observar que los clientes que cometieron fraude, tienden a reportar un número de vehiculos involucrados de 1 o 3 carros, con lo que se ve que más de una persona se encuentra involucrada en este tipo de actos. Por otro lado, se ve que los casos que más datos tienen, tanto de fraude como no fraude es 1 o 3 vehiculos involucrados, por lo que puede ser un comportamiento aprendido por los defraudadores. Por otro lado, los menores casos se presentan cuando hay 4 o 2 vehículos involucrados.

La variable ocupación del asegurado, parece tener una incidencia o correlación media con el reporte de fraude, pues se observa que la profesión de executive managerial tiene una alta indice de fraude llegando a ser el 37.5% de los casos en los que es reportó fraude. El resto de categorias, parece tener una incidencia baja sobre el fraude, pues mantienen proporciones similares con no fraude a través de la gráfica



Al revisar la variable tipo de incidente, se ve que esta tiene gran correlación con la comisión de fraude, observandose que los clientes que cometieron fraude, tienden a hacer el fraude con incidente de multiple colision de vehiculos y colisión con un solo vehículo, se ve que sus proporciones son de aproximadamente el 26% de los casos. en contraste menos del 10% de los casos reportados como fraude corresponden a robo de vehículo o incidente al encontrarse parqueado.

La variable estado donde ocurrió el incidente es una variable que ayuda a determinar en una pequeña medida si existió fraude o no para este dataset. Puntualmente, en la categoría de Ohio se observa una alta proporción de personas que cometieron fraude en comparación con las que no, siendo cerca del 45% vs el 55% correspondientemente. Por otro lado, se observa que el estado en donde más casos de fraude se cometieron, es el estado de South Carolina en donde cercan de 75 de las reclamaciones fueron fraudulentas

**Preprocesamiento de la data y entrenamiento del modelo**

Como el objetivo de nuestro proyecto, mencionado anteriormente, es la clasificación de casos fraudulentos y la generalización de los resultados en datos no observados, es preciso eliminar variables que no aportarán en la consecución de este objetivo como lo son insured zip pues este varía de cliente a cliente dado que se encuentra atado a la residencia de los mismos, además que podría generar problemas de discriminación por parte del modelo, se elimina la variable policy\_number pues es única por cliente, incident\_location ya que esta se asocia a la ubicación donde ocurre el siniestro y por lógica no debería tener incidencia o correlación dentro de la incurrencia en fraude, y las variables auto\_model y policy\_bind\_date pues su información es redundante con las variables auto\_make y months\_as\_customer respectivamente.

Al eliminar las variables anteriormente mencionadas, nuestro dataset; el mismo se reduce a 29 variables o features, una variable label y 1000 registros. Las cuales corresponden a 1 variable tipo flotante, 15 variables tipo entero, y 17 variables tipo objeto

Al hacer la revisión de los datos faltantes en este dataset, pareciera que no existen variables que tengan algún registro faltante, sin embargo, se observó que las variables property\_damage, police\_report\_available, collision\_type se les había anotado el simbolo (?) previamente para llenar los espacios correspondientes a los datos faltantes, esto fueron visualizados en nuestro análisis de data, y fueron modificados nuevamente para indicarle al modelo que estos datos correspondían a missing data, esto porque el algoritmo que se utilizó XGBoost, no tiene problema con los datos faltantes en la construcción del modelo. A pesar de lo anterior, y para evitar mejorar los resultados de nuestro modelo, se realizó la imputación de los valores faltantes mediante la estrategia de relleno con base en los valores más frecuentes.

Se realizó la codificación de las variables categóricas y de la label. Para esta última se utilizó como representación valor de 1 para las reclamaciones ilegítimas y 0 para las reclamaciones legítimas. Por su parte, para la codificación de las variables categóricas se utilizó One-Hot-Encoding, el cual asigna un código/valor a cada uno de los valores que tiene estas variables.

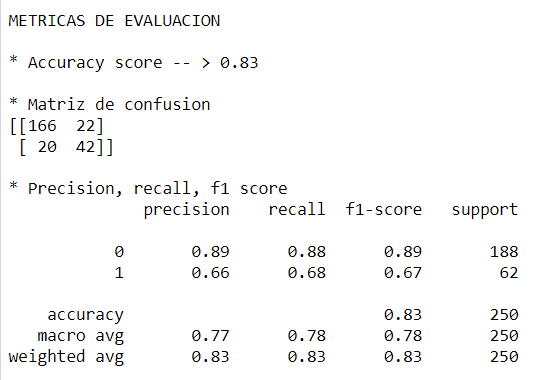
De igual manera, se utilizó stratify como método para lidiar con el problema de imbalance data. Para ello; stratify mantiene las mismas proporciones que se encuentra en el dataset original de datos fraudulentos como no fraudulentos en la selección de su muestra para los datos de entrenamiento, test y validación.

Finalmente cabe resaltar, que para contrarrestar a los problemas de overfitting que trae consigo el uso del modelo de XGBoost, fue necesario tomar una muestra de los datos del 90% y de las columnas del 80% en su entrenamiento. Estos valores se fueron ajustando y se realizó la iteración con los mismos 3 veces, esto porque valores de muestra del 50 – 80%, y valores donde las columnas eran de 50 -70% presentaron métricas de clasificación muy bajas, haciendo perder la eficacia del modelo XGBoost.

**Métricas de clasificación**

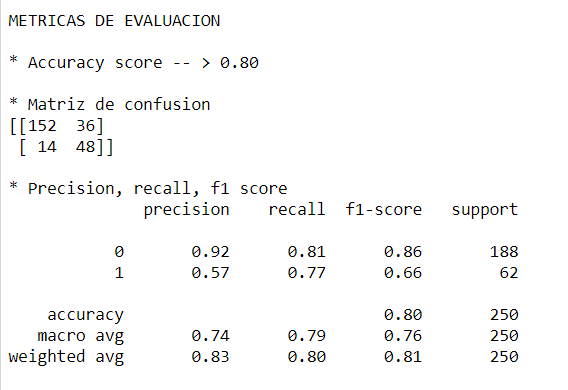
Debido a que estamos trabajando con un problema de imbalance data, la métrica de accuracy/ precisión no es la medida más adecuada para evaluar los resultados de desempeño de nuestro modelo. Esto porque altos niveles de accuracy pueden ser alcanzados por modelos en los que se logre clasificar la mayoría o la totalidad de los datos de la clase que más representantes tiene en el dataset y mis-clasificar o clasificar incorrectamente la totalidad de los datos que menos casos tiene en el dataset. Por ende, y dado que el objetivo de nuestro proyecto es encontrar un modelo que clasifique correctamente la clase fraude reportado; la medida de desempeño de nuestro modelo no puede ser accuracy.

Para nuestro caso particular, se utilizarán las medidas de precision, recall y f1 – score porque contribuyen en mejor medida a la evaluación de desempeño del modelo, siendo precisión el porcentaje de los valores clasificados como positivos correctos, para ello se utiliza la formular True positives dividido por la cantidad de True positives y false positives, queriendo decir, que se revisarán los valores que el modelo clasifique como positivo o fraudulento, y se verificará la cantidad de valores en los cuales la predicción correspondió al valor verdadero. Por otro lado la métrica recall, nos muestra una relación en donde podemos observar la cantidad de observaciones que se tuvieron correctas en una clase, dividido por la cantidad de observaciones que pertenecían a dicha clase. De igual manera, y dado el trade-off que tienen estas dos métricas donde el incremento de una representa el detrimento de la otra, se utilizará la métrica de f1 – score, la cual es sencillamente el valor de la media armónica entre estas dos métricas, precisión y recall.

Los resultados se muestran a continuación, representan los valores para dichas métricas. Cabe resaltar que estos resultados se obtuvieron con la configuración de los parámetros como los trae por default XGboost, y que para su entrenamiento se utilizó el 90% de la data y el 80% de las columnas para evitar problemas de overfitting.

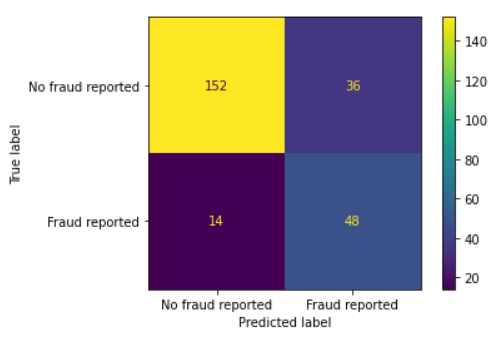
Como se observa, el modelo tiene un accuracy del 83%, sin embargo, se deben entrar a revisar otras métricas para verificar efectivamente el modelo está clasificando correctamente los valores para nuestra variable de interés. Para este caso, en donde fraude es representado por el valor 1, se tiene medidas de desempeño de precisión del 66%, recall del 68% y f1 – score del 67%, los cuales son buenos indicadores, si tomamos en cuenta que en el caso colombiano solamente el 7 - 8% de los casos de fraude se logra identificar.

Para ir un poco más lejos e intentar mejorar las métricas de desempeño, se realizó el tuneo de algunos de los parámetros que trae por defecto XGBoost. La optimización de dichos parámetros se realizó mediante la función de GridSearch() la cual encuentra los mejores parámetros para ser ajustados a nuestro caso, con la desventaja de que es un proceso que consume bastantes recursos computacionales al igual que tiempo, por lo cual se realizó un promedio de 15 iteraciones, 2 de las cuales se dejaron documentadas, pues son las que mejores resultados arrojaron. Igual que en el caso anterior, para evitar problemas de overfitting, se solicitó al algoritmo tomar el 90% de la data y el 80% de las columnas para ajustar el modelo. Los resultados del mismo se ven a continuación.



A pesar de que después de esta optimización nuestro modelo redujo su metrica de accuracy del 82% al 80%, la cual como se mencionó anteriormente no es la mejor medida para nuestro caso; se observa una mejoría notable en los valores de recall para nuestra clase de interés, fraude igual a 1, donde esta medida pasó de ser del 68% a ser del 77%. En términos prácticos, esto representa que una mayor cantidad de los casos de fraude están siendo predichos y clasificados como fraudulentos.

Lo anterior se observa más fácilmente en la matriz de confusión, donde logra apreciar que los valores que habían sido predichos como fraude y que realmente correspondían a fraude, pasaron de 42 a 48 mediante esta optimización, se veía pues que anteriormente, se escapaban 6 casos fraudulentos, los cuales eran clasificados como no fraude, todo esto significa que el modelo está realizando una mejor clasificación para nuestra clase de interés.

** Modelo con parámetros por default Modelo con parámetros optimizados**

**Conclusiones**

Mediante este proyecto se logra apreciar la fuerza y eficacia que tienen las herramientas de machine learning en la actualidad y da una ilusión del impacto positivo que podrían llegar a tener en el ámbito colombiano. Colombia es uno de los países donde más se ha normalizado la comisión fraude debido a la falta de investigación y castigo de este tipo de hechos, como se menciona en una investigación hecha por el tiempo, las compañías aseguradoras tienen indicio de fraude en 7 de cada 10 casos, sin embargo, solamente entran a confirmar este hecho en 7 – 8 % de los casos. Esto se da en gran medida debido a que los costos de investigación de estos hechos, serían superiores a los que se debe pagar por dicha póliza, lo que lleva a su vez a la normalización de estos casos al dejar en la impunidad al defraudador. Por medio del modelo de XGBoost, se logró apreciar que se pueden llegar a tener métricas de recall del 77%, indicando que se podría llegar a identificar el 77% de los casos de personas que cometen este hecho y se podría enfocar los recursos de la compañía en dichas investigaciones.

Por otra parte, gracias a este proyecto, se dejaron las bases sentadas para una investigación en el ámbito colombiano cuando se cuente con la data necesaria para alimentar este modelo, y se puede dejar como recomendación empezar a incluir dentro de los datos que recolectan las compañías aseguradoras los hobbies del asegurado, los montos discriminados por tipo de daño, la severidad del incidente entre otros, pues parecen tener una alta correlación con la comisión de fraude.

**Referencias bibliograficas:**

https://dev.to/grayhat/exploratory-data-analysis-using-python-28h

https://towardsdatascience.com/for-real-auto-insurance-fraud-claim-detection-with-machine-learning-efcf957b38f3

https://www.youtube.com/watch?v=GrJP9FLV3FE&list=PLblh5JKOoLULU0irPgs1SnKO6wqVjKUsQ&index=8

https://towardsdatascience.com/beginners-guide-to-xgboost-for-classification-problems-50f75aac5390

https://evidenciasenpediatria.es/files/41-13481-RUTA/Fundamentos\_13.pdf

https://www.eltiempo.com/economia/sectores/el-70-de-los-reclamos-a-las-aseguradoras-tienen-indicio-de-fraude-406712

https://www.siras.com.co/siras/NoAutenticado/Soat

file:///C:/Users/afinf/Downloads/405-Texto%20del%20art%C3%ADculo-766-1-10-20180912.pdf

file:///C:/Users/afinf/Downloads/629-Texto%20del%20art%C3%ADculo-1115-1-10-20200727.pdf

https://inif.com.co/blog/2018/07/23/clonacion-tarjetas-de-credito/

https://repository.usta.edu.co/bitstream/handle/11634/30529/2020helmervargas.pdf?sequence=2&isAllowed=y

https://revistas.ucr.ac.cr/index.php/ingenieria/article/view/6431/6136

https://www.jeremyjordan.me/imbalanced-data/

https://towardsdatascience.com/beginners-guide-to-xgboost-for-classification-problems-50f75aac5390