Travel Insurance

Proyecto final de Data Science para CoderHouse

NOVIEMBRE 2021

Autor: Leandro Hornos

EMAIL: leandrohornos@gmail.com
GITHUB: github.com/leanderShulgin
LINKEDIN: linkedin.com/in/leandro-adriel-hornos/

Introducción

El cliente es una empresa de viajes y excursiones que ofrece un paquete de seguro de viaje a sus clientes. La empresa requiere saber qué clientes estarían interesados en comprarlo en función del historial de su base de datos. El seguro se ofreció a algunos de los clientes en 2019 y los datos proporcionados se han extraído de las ventas del paquete durante ese período. Los datos se proporcionan para alrededor de 2000 de sus clientes

Objetivo

Se desea conocer el perfil de cliente que está interesado en la adquisición del seguro de viajes. Además, la empresa desea contar con un modelo de machine learning capaz de identificar a los potenciales compradores del seguro. Finalmente, la empresa está considerando crear una cobertura especial para covid-19 y desea saber si es posible inferir el interés respecto a este producto a partir de los datos

Exploración del dataset

Variables

- **Age**- Edad del cliente
- **Employment Type**: Sector en el que trabaja el cliente (público o privado)
- **GraduateOrNot**: Si el cliente se graduó o no de la universidad
- **AnnualIncome**- Ingreso anual del cliente (redondeado a las 50k rupias más cercanas)

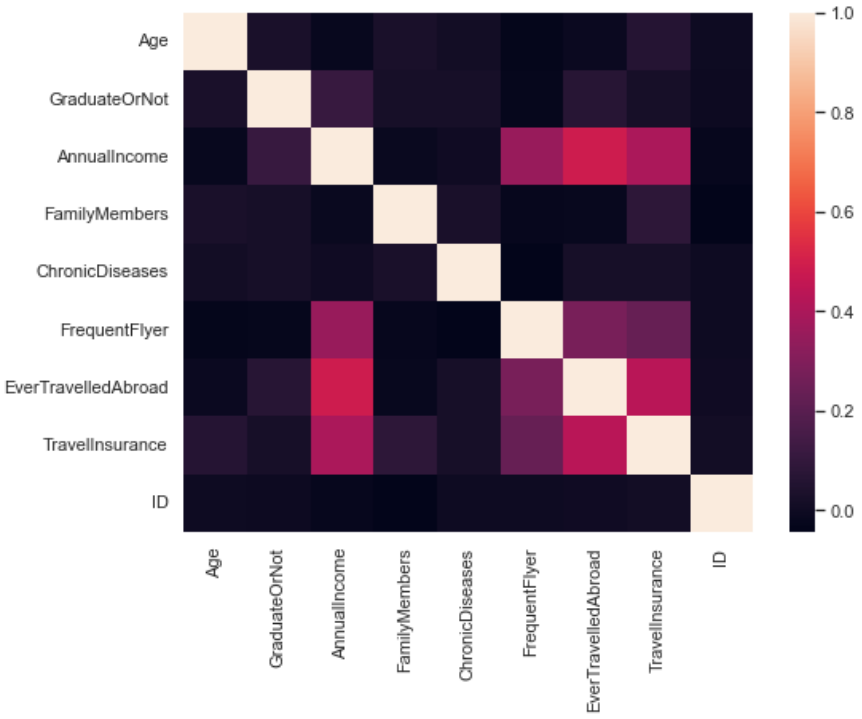
- **FamilyMembers**: Número de miembros en la familia del cliente
- **ChronicDisease**: Si el cliente tiene alguna enferemedad crónica
- **FrequentFlyer**: Si adquirió un paquete de viaje en 4 más ocasiones en los dos últimos años
- **EverTravelledAbroad**: Si el cliente ha viajado alguna vez al exterior, sea con esta compañía u otra distinta.
- **TravelInsurance**: Si el cliente adquirió el paquete de seguro ofrecido por primera vez en 2019

Preparación de los datos

La exploración del dataset muestra que no existen valores nulos o datos evidentemente erróneos. La mayoría de las variables son binarias pero están expresadas de distintas formas (si y no, 0 y 1, etc). Se unifican todas a variables binarias que valen 0 o 1.

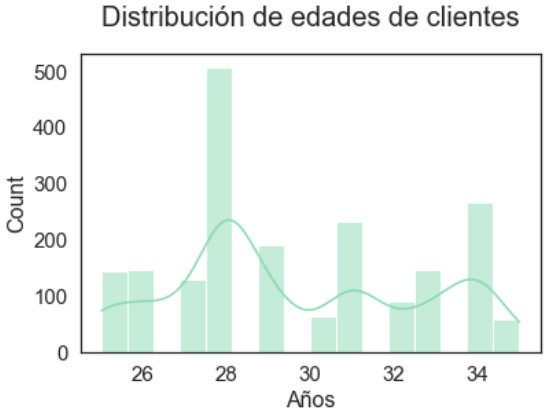
Correlaciones entre variables

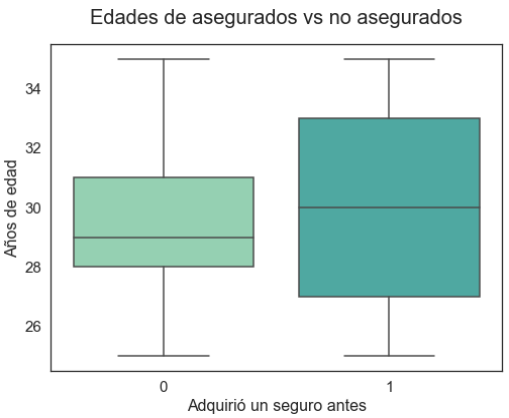
En un heatmap de correlaciones se puede ver que parece cierta correlación entre el ingreso de los clientes y el hecho de ser viajeros frecuentes, haber viajado al exterior anteriormente y haber adquirido el seguro. Concretamente en relación a la adquisición de un seguro, se ve que la correlación más fuerte es con el hecho de haber viajado anteriormente al exterior (EverTravelledAbroad).



Edad

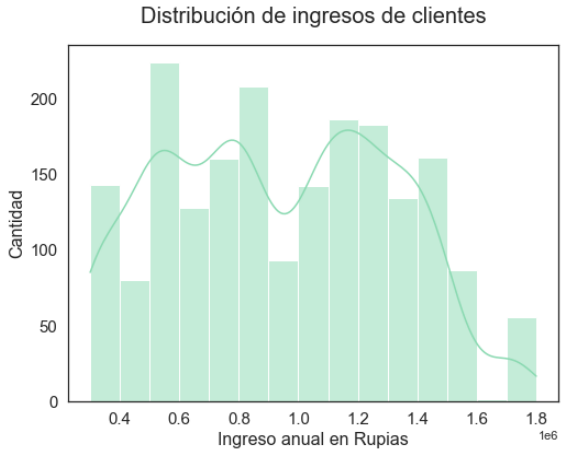
Los clientes son personas jóvenes entre los 25 y 35 años. Dentro de estas edades la distribución es bastante pareja, aunque con cierta asimetría hacia los más jóvenes, con una edad media cercana a los 29 años.
Puede observarse que la distribución de edades entre quienes adquieren un seguro de viaje y quienes no es bastante similar. Se nota más concentración de edad alrededor de la media entre quienes no adquieren seguros de viaje. Más allá de leves tendencias, no se observa que la edad sea un factor determinante en la decisión de adquirir un seguro.

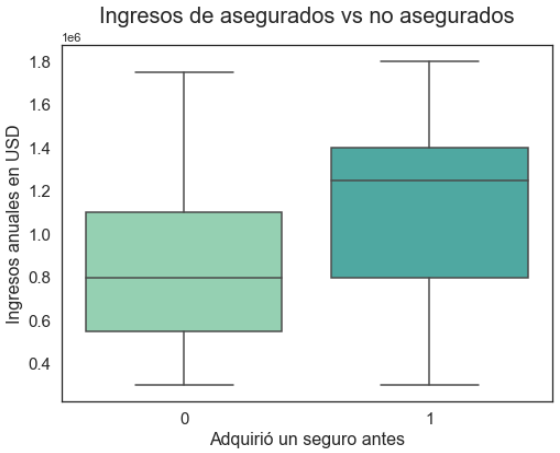




Posición económica

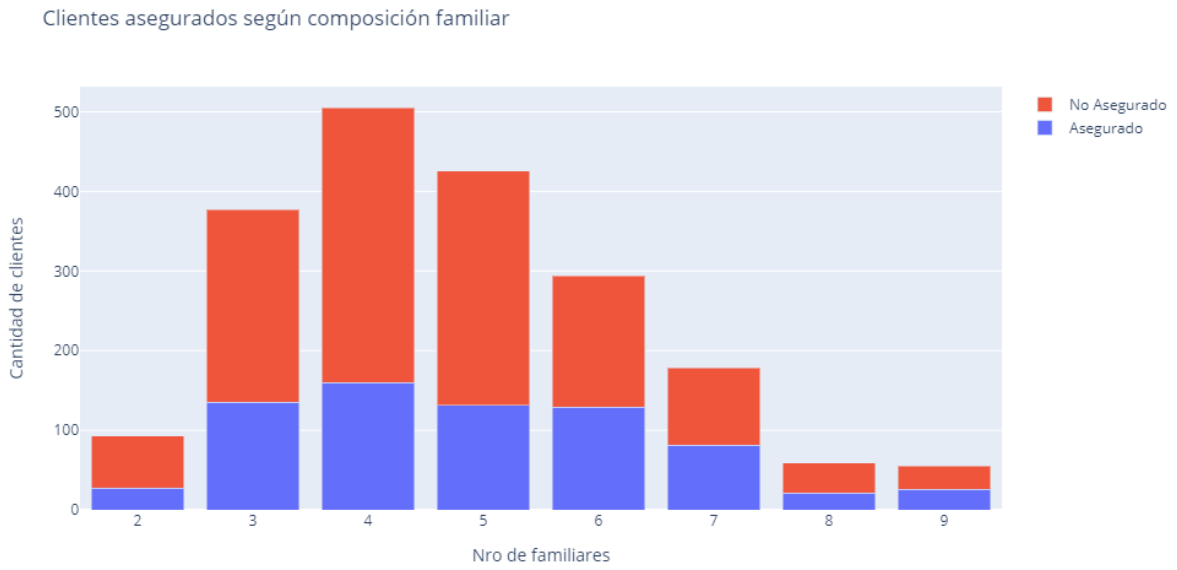
Los ingresos de los clientes se distribuyen en un rango bastante amplio, pero la distribución es bastante homogénea para lo que suele ser una distribución de ingresos, donde la mayor cantidad de individuos suele agruparse en los ingresos más bajos.
Si bien existe bastante solapamiento entre ambos grupos, puede verse que las personas que adquirieron un seguro tienden a tener mayores ingresos que aquellas que no.





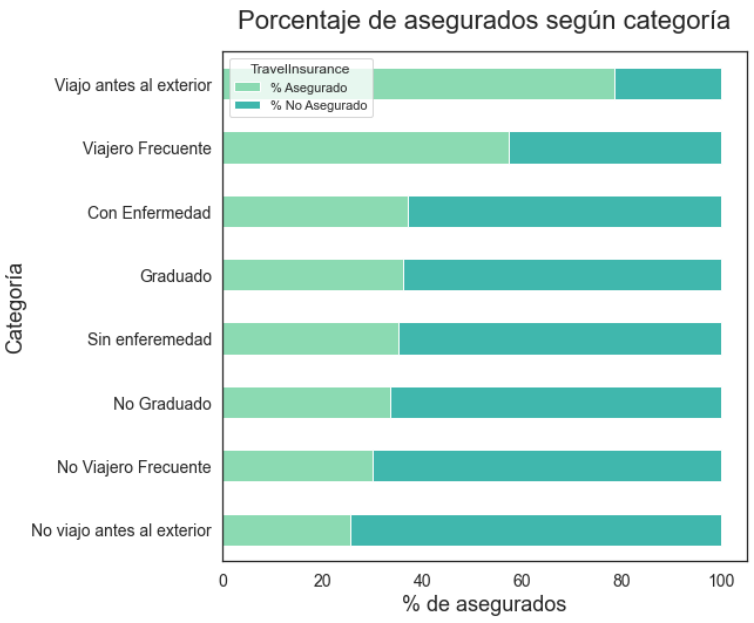
Composición familiar

La mayoría de los clientes tienen familias de unos 3 a 5 miembros. La proporción de asegurados es baja cuando són solo dos miembros, mientras que es más alta en familias de tamaño intermedio



Variables binarias

El dataset contiene un conjunto de características de los clientes definidas como variables binarias, las cuales son enfermedad crónica, viajero frecuente, viajes al exterior en el pasado y si se han graduado. Podemos estudiar estas variables comparativamente, viendo como es la proporción dentro de cada subgrupo, según si poseen o no alguna de dichas características.



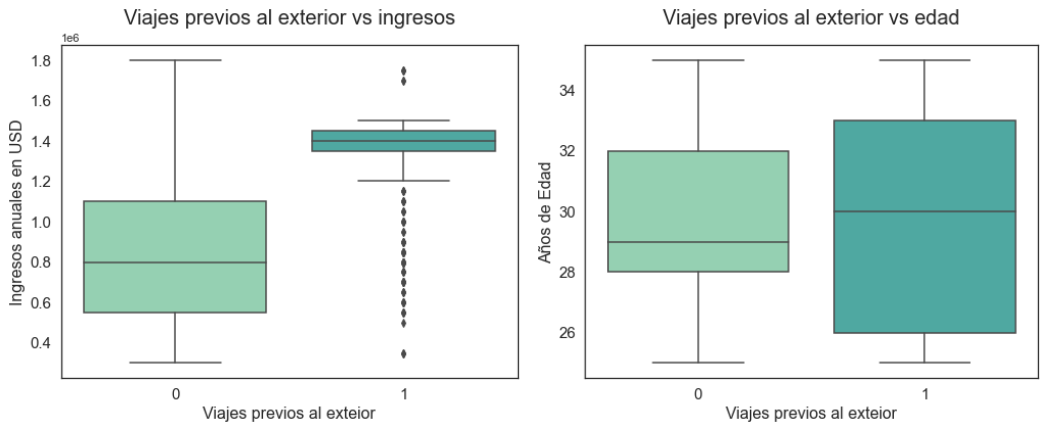
Se observa en el gráfico que el mayor porcentaje de asegurados se encuentra entre quienes han viajado antes al exterior. Además, quienes no hicieron viajes al exterior son los que menos contrataron seguros, por lo que esta característica es la que muestra la mayor disparidad a la hora de elegir o no contratar un seguro. En segundo lugar se encuentran los viajeros frecuentes cómo los que más seguros contratan, habiendo también una gran diferencia con los que no lo son. Es posible que exista un alto solapamiento entre los viajeros frecuentes y quienes han viajado al exterior anteriormente. Habría que ver cuan es lo que lleva a este grupo a adquirir seguros, puede ser que malas experiencias en viajes

anteriores los haya vuelto más precavidos, o bien que al viajar al exterior es más probable que deseen contratar un seguro ya que en dicha situación no tienen acceso a los recursos que si encuentran en su país de origen.

Relación entre viajes previos al exterior y otras variables

Respecto al ingreso, en el heatmap de correlaciones se puede apreciar que una de las correlaciones más altas de la variable EverTravelledAbroad se da con AnnualIncome, es decir, con el poder adquisitivo de los clientes. En los boxplots que aquí se muestran se observa que quienes viajaron antes al exterior tienen un salario medio claramente superior a quienes no lo han hecho. Además, el rango de ingresos de quienes viajaron antes al exterior está muy concentrados alrededor de la mediana, es decir, la mayoría tiene salarios altos. Se observan muchos "outliers" hacia menores ingresos en quienes han viajado al exterior, que podría deberse a que dichas personas viajaron por motivos laborales y no pagaron los gastos del viaje de su propio bolsillo, o bien que en alguna ocasión han hecho un esfuerzo económico excepcional para hacerlo.

Respecto a la edad de quienes viajaron antes al exterior, podemos ver es ligeramente mayor que quienes no, lo cual puede estar correlacionado con mayores ingresos. Sin embargo, hay una mayor dispersión en la edad entre quienes viajaron antes al exterior. No parece haber mucha relación entre la edad y dicha variable.



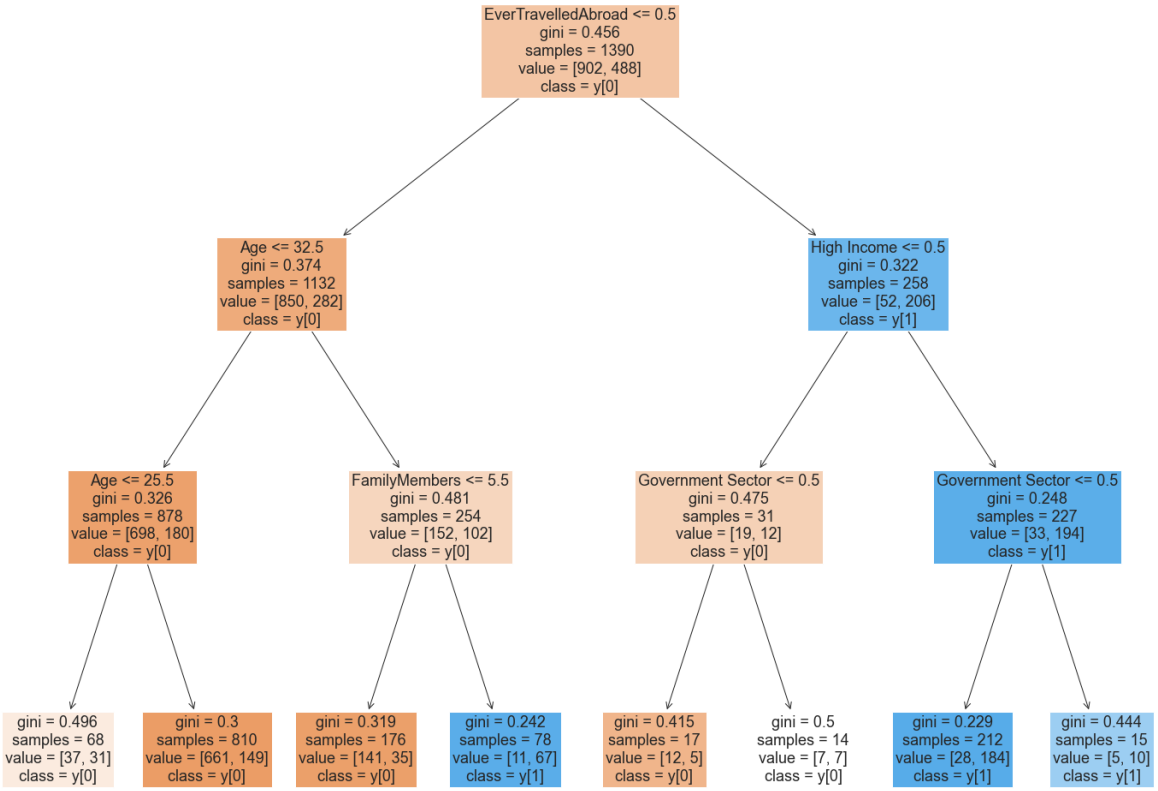
Modelos

Por lo que observamos en el dataset, la mayoría de las variables son categóricas, y dentro de ellas muchas son binarias, incluyendo la variable a predecir: TravelInsurance. Es por ello que un árbol de decisiones surge como un modelo atractivo para intentar predecir si un cliente adquirirá un seguro.

Vimos en el análisis de las variables binarias que aquella que parece tener más peso en la contratación de un seguro es EverTravelledAbroad. Vamos ahora a utilizar un árbol de decisiones para ver si somos capaces de modelar la variable TravelInsurance a partir de las variables del dataset.

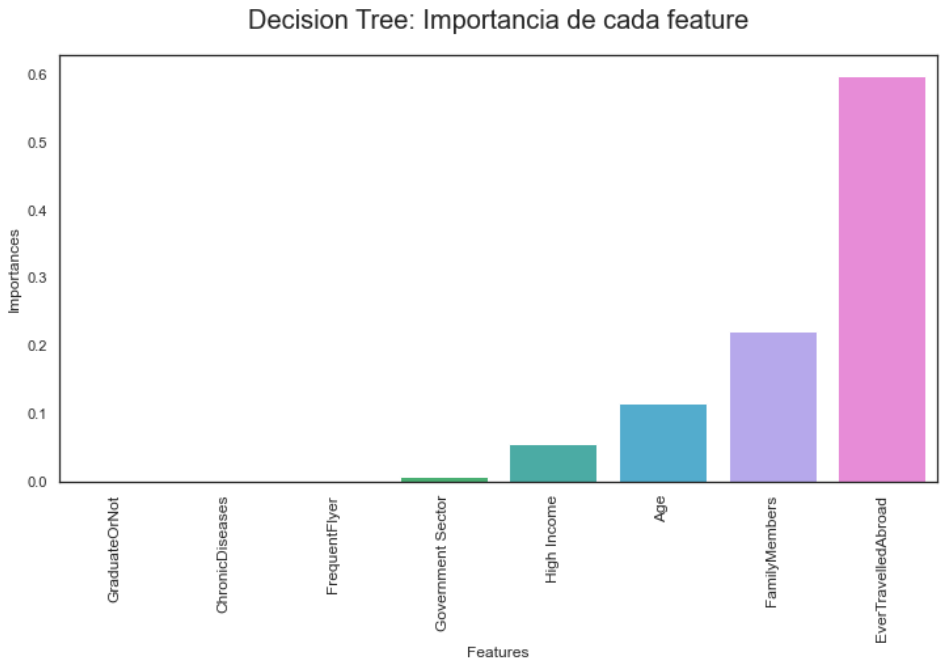
Modelo 1: Árbol de decisión

Para encontrar los mejores hiperparámetros para el modelo se utilizó GridSearchCV de SKlearn y se obtuvo que para el árbol óptimo la mejor profundidad es 3 y la mejor cantidad minima de muestras por hoja es 10.
Se ajustó el modelo con los hiperparámetros optimizados y se obtuvo la siguiente estructura



Vemos que, como esperábamos a partir del análisis exploratorio univariado, la característica más importante es EverTravelledAbroad, es decir, si el cliente ha viajado antes al exterior. Vemos que le siguen en menor medida la composición familiar y la edad, variables que habíamos observado parecían tener alguna influencia. Por otro lado, si bien vimos que si un cliente además de haber viajado al exterior entraba en la categoría "FrequentFlyer" tenía más posibilidad de haber adquirido un seguro. Sin embargo, la cantidad de clientes que caen en ambas categorías es muy pequeña y puede verse que en el arbol de decisiones no tiene importancia.

Es interesante notar también que la variable ChronicDiseases no parece tener ninguna influencia en la decisión de adquirir un seguro, esto es importante ya que la empresa desea ofrecer un seguro contra el Covid. Si bien en el pasado el poseer enfermedades crónicas no fue relevante, no puede saberse de estos datos si la pandemia actual no afectará a esta variable ya que alguien con enfermedades crónicas puede contraer una forma más grave de covid.



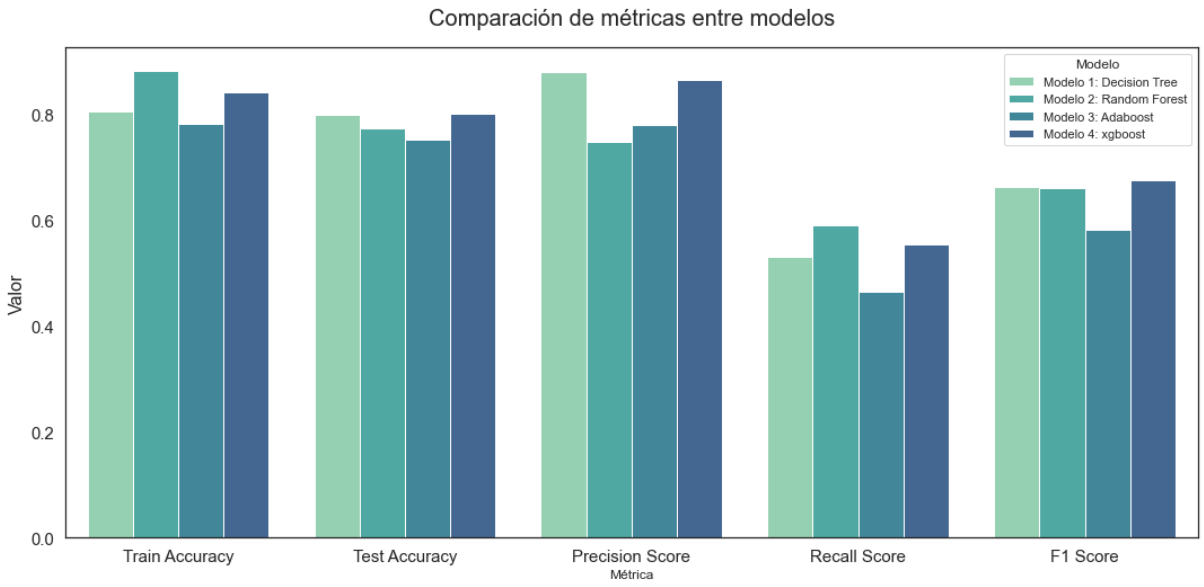
MODELO DE ARBOL DE DECISON			
Accuracy	**Precision**	**Recall**	**F1**
0,80	**0,88**	**0,53**	**0,66**

Podemos ver en las métricas que el modelo tiene un buenos accuracy y precision scores. En cambio el recall es bastante bajo. Esto no es lo ideal ya que queremos maximizar la identificación de casos negativos.

Se prueban modelos más sofisticados: random forest y modelos de boosting: adaboost y xgboost

Otros modelos

Se comienza probando un random forest, utilizando GridSearchCV para encontrar los hiperparámetros óptimos. Además, se prueban dos modelos de boosting, adaboost y xgboost. Se calculan las métricas de cada modelo y se comparan en el siguiente gráfico



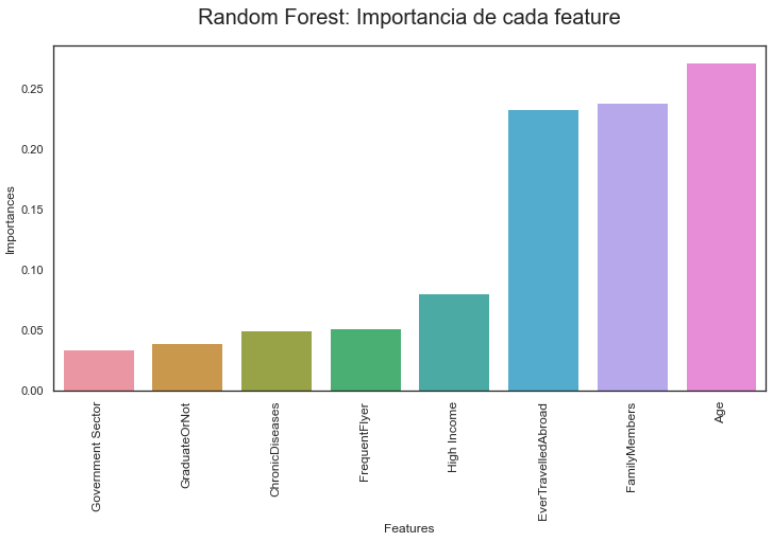
Se puede ver que no se ha logrado una gran mejora con Adaboost respecto a lo que ya se tenía. Sin embargo, con XGboost vemos que se logra un recall bastante cercano al del random forest, pero con una precisión mucho mayor, lo que se traduce en que este modelo sea el que tenga el mayor F1 score de todos.

Buscando un equilibrio entre la complejidad del modelo y sus métricas, se concluye que **el modelo de random forest es el más adecuado**, ya que es el que tiene el deja la menor cantidad de potenciales compradores de lado (mejor recall) y da importancia a las características que vimos que se destacan entre los compradores.

Por último, si bien el random forest tiene menor precisión que sus alternativas, el porcentaje de falsos negativos es de alrededor de 25%, lo cual resulta bastante aceptable si se tiene en cuenta que dichos falsos negativos comparten muchas características con quienes sí compraron el seguro, por lo que sería interesante incluirlos en la oferta ya que existe la posibilidad de que sí compren el seguro la próxima vez.

Modelo 2: Random Forest

Aquí se muestran las importancias que da a cada feature el modelo de random forest. Vemos que sigue dando alta relevancia a la variable EverTravelledAbroad pero ahora crecen en importancia la edad y la composición familiar con respecto a lo que se había obtenido con el árbol de decisión.



Aplicabilidad económica del modelo

Se probaron distintos modelos de clasificación a fin de identificar correctamente a aquellos clientes con mayor potencial de adquirir un seguro de viaje.
De los modelos elegidos, el árbol de decisiones sencillo es el que da la pauta más clara a la hora de clasificar a los clientes, y se vió que un modelo más sofisticado, de Random Forest, da mejores métricas y por lo tanto fue el elegido.

Se vió también que el modelo genera pocos falsos positivos pero sí una alta proporción de falsos negativos. De hecho, se eligió el modelo de random forest ya que la tasa de falsos negativos era un poco mejor que en los demás modelos, a expensas de convivir con una tasa un poco más alta de falsos positivos.

Existen dos alternativas posibles:

1. ofrecer el producto a todos los clientes
2. aplicar el modelo y ofrecer el producto sólo a aquellos clientes clasificados como potenciales candidatos

En el primer caso, estamos maximizando las ventas ya que el ofrecimiento llegará a todos los clientes de la compañia. En el segundo caso, estaremos minimizando las pérdidas por ofrecer el producto a clientes que no tienen interés, ya que el modelo es muy confiable prediciendo quien estará interesado el en seguro, pero pasa por alto muchos potenciales clientes.

La elección de una u otra estrategia dependerá básicamente del costo de promoción del producto. Si el mismo es despreciable, conviene definitivamente ofrecérselo a todo el mundo. Por otro lado, si el costo es muy elevado (envío de merchandising, ventas personalizadas, etc), entonces quizás sea conveniente aplicar el modelo y minimizar pérdidas.

Para ilustrar esto, consideremos la diferencia entre las ganancias generadas aplicando o no el modelo.

Ganancia=Ingresos por ventas−Costo de promoción

Supongamos que V es el dinero que se gana por cada seguro vendido, sin tener en cuenta el costo de promoción.

Llamemos X al costo de promoción del producto y definamos a X como una fracción de V, que va entre 0 y 1. O sea, X da la proporción de la venta que se pierde por promocionar el producto; si X=0 la promoción es gratis, si X=1 se gasta la misma cantidad de dinero el promocionar el producto que lo que se gana en la venta, con lo cual la ganancia neta es cero.

Llamamos N_T al número total clientes y N_M a los clientes identificados como positivos por el modelo.

Finalmente, llamamos P a la proporción de compradores entre los clientes y P_M a los verdaderos compradores dentro de los identificados por el modelo.

o sea que la diferencia de ganacias es:

$$Ganancia_{STD} = N_T PV - N_T XV$$

$$Ganancia_{Modelo} = N_M P_M V - N_M XV$$

$$Ganancia_{Modelo} - Ganancia_{STD} = N_M P_M V - N_M XV - N_T PV + N_T XV$$

reagrupando nos queda

$$Ganancia_{Modelo} - Ganancia_{STD} = V(N_M P_M - N_T P) + XV(N_T - N_M)$$

La ganancia aplicando el modelo será mayor a la de no aplicarlo cuandol la expresión del lado derecho de la ecuación sea mayor que cero. Cancelamos V entonces y nos queda que

$$Ganancia_{Modelo} > Ganancia_{STD}$$

si

$$(N_M P_M - N_T P) + X(N_T - N_M) > 0$$

N_T-N_M es un valor que siempre es mayor que cero, ya que el primero es el número total de clientes y el segundo en el número de potenciales compradores identificados por el modelo. Por lo tanto, a medida que aumenta X crece la ventaja de utilizar el modelo frente a ofrecer el paquete a todos los clientes, ya que se pierde menos dinero en ofrecimientos que no se traducen en ventas.

Por otro lado, si X=0, es decir, si ofrecer el paquete es gratis, entonces el segundo término se anula y nos queda $N_M P_M$-$N_T P$, siendo el primer término los verdaderos positivos identificados por el modelo, mientras que $N_T P$ son todos los positivos. Como los casos identificados por el modelo van a ser siempre menores o iguales que el total, dicho término es siempe menor o igual que cero. Esto significa que, *si el costo de ofrecer el producto es despreciable entonces se pierde dinero al aplicar el modelo, ya que el mismo deja afuera casos positivos.

Recordemos que P_M es la proporción de verdaderos compradores dentro de los identificados positivamente por el modelo, o sea *P_M es la precisión del modelo*.

$$P_M = Precision = \frac{Verdaderospositivos}{verdaderospositivos + falsospositivos}$$

Por otro lado,N_TP son todos los compradores, es decir, los *verdaderos positivos + falsos negativos*. O sea que podemos relacionar N_TP con el Recall como:

$$Recall = \frac{Verdaderospositivos}{verdaderospositivos + falsosnegativos} = \frac{N_M P_M}{N_T P}$$

Usamos estas dos ecuaciones para reemplazar los valores de N_M y P_M en la ecuación anterior para llegar a una expresión que solo depende de la calidad del modelo y las características del dataset. Por lo tanto, tenemos que

$$Ganancia_{Modelo} > Ganancia_{STD}$$

si

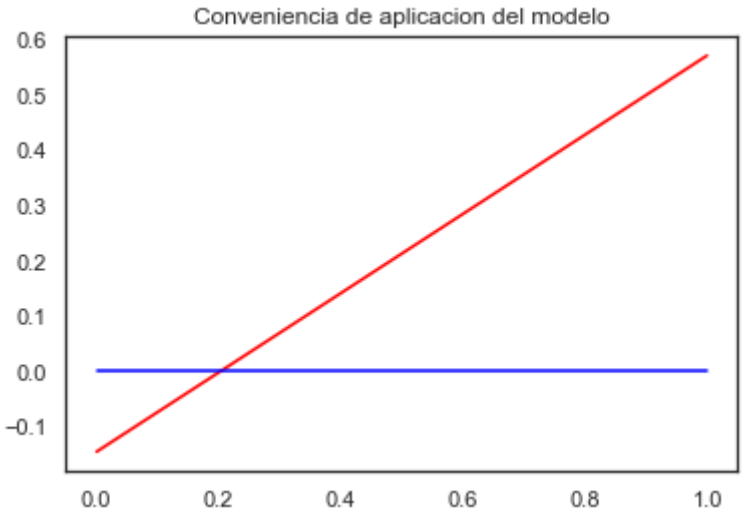
$$P(Recall - 1) + X(1 - P\frac{Recall}{Precision}) > 0$$

El lado izquierdo de la desigualdad corresponde a la ecuación de una recta en función de X. Podemos graficar la recta para ver en que punto pasa por cero. A partir de dicho valor de X aplicar el modelo supondrá mayores ganancias que ofrecer el producto a todos los clientes.

Para Random Forest tenemos:

- P = 0.357
- Recall = 0.59
- Precision = 0.749

Graficando la recta se obtiene:



Vemos, como esperábamos, que para costos bajos la diferencia de ganancias es negativa, pero se hace positiva aproximadamente a partir de 0,2. Es decir: **conviene aplicar el modelo para identificar potenciales clientes cuando el costo de la promoción del producto supera en aproximadamente un 20% la ganancia obtenida por la venta**

Conclusiones

- Se estudiaron las características de los clientes que adquieren el seguro y se detectó que aquellas personas que han viajado antes al exterior o son viajeros frecuentes, que tienen una familia de 4 o más miembros y que tienen mejor posición económica son los mejores candidatos a la hora de ofrecer un seguro de viaje.
- Se desarrolló un modelo de Machine Learning (Random Forest) capaz de clasificar a los clientes e identificar aquellos con alto potencial de adquirir el seguro
- El modelo desarrollado es mucho más eficiente minimizando las pérdidas que maximizando las ventas, ya que genera pocos falsos positivos pero pasa por alto una gran proporción de los potenciales compradores
- Se mostró que el modelo es económicamente ventajoso cuando el costo de promoción del seguro es una fracción elevada de la ganancia obtenida por la venta, y se estimó que cuando dicho costo supera el 20% es conveniente aplicar el modelo
- Respecto a un potencial interés frente a un seguro de cobertura por COVID-19, vemos que el dataset no cuenta con información relevante al respecto como para poder hacer alguna clase de producción. La única variable relacionada con la salud, la presencia de enfermedades crónicas, no tuvo ningún peso a la hora de decidir la adquisición del seguro. Sin embargo, dado que el seguro fue ofrecido antes de la pandemia es probable que dicha variable tenga influencia en futuras compras