

Trabajo Práctico Final - Grupo 1

Alumnos

- Jacinto, Milagros Aylén - 15938/5
- Loza Bonora, Leonardo Germán - 16181/7
- Moschettoni, Martín - 15836/0

Alumnos

Descripción del conjunto de datos elegido

Introducción

Explicación del dominio

Forma de recolección del conjunto de datos

Infomación sobre el conjunto de datos

Preguntas de interés

Hipótesis

Preprocesamiento de datos

Experimentos realizados

Agrupamiento

Análisis realizado

Resumen

Conclusión

Clasificación

Modelos construidos

Naive Bayes

Árbol de Decisión

Árbol de decisión sin campañas anteriores

Red Neuronal

Comparación entre modelos

Conclusión

Conclusión Final

Referencias

Descripción del conjunto de datos elegido

Introducción

Analizar la personalidad de los clientes sirve para que las empresas puedan modificar sus productos en función de sus clientes objetivo. Por ejemplo, en vez de gastar dinero para vender un producto general para todos los clientes, se puede analizar qué grupo de clientes tiene más probabilidad de comprar un producto específico y comercializarlo con ese sector.

Para lograr este trabajo es necesario que podamos responder varias preguntas como: ¿Qué dice la gente de los productos? ¿Qué hacen nuestros clientes?. Esto se conseguirá utilizando diferentes técnicas de agrupamiento que nos muestren diferentes patrones y nos brinden información relevante de los mismos.

Una vez analizados los perfiles de clientes, se intentará generar modelos que puedan predecir si los clientes aceptarán una campaña.

Explicación del dominio

iFood es la aplicación de entrega de alimentos líderes en Brasil, posee el 80% del mercado. Vende productos de 5 categorías principales: vinos, carnes raras, frutas exóticas, pescados especialmente preparados y productos dulces. Estos se puede dividir además en oro y productos regulares. Los clientes pueden ordenar y adquirir productos a través de 3 canales de venta: tiendas físicas, catálogos y sitios web de la empresa.

A nivel mundial, la compañía tuvo ingresos sólidos y un resultado final saludable en los últimos 3 años, pero las perspectivas de crecimiento de las ganancias para los próximos 3 años no son prometedoras. Por esta razón, se están considerando varias iniciativas estratégicas para revertir la situación. Una de ellas es mejorar el rendimiento de las actividades de marketing, con un enfoque especial en las campañas.

A raíz de esta situación la empresa iFood generó un puesto de trabajo nuevo con el objetivo de construir un modelo predictivo que produzca mayor ganancia para la próxima campaña de marketing directo. Para ello presentan un conjunto de datos con metainformación sobre los clientes y sobre las interacciones de la campaña iFood con cada cliente, donde se busca encontrar características sobre los clientes que luego se reflejen en el conjunto de datos de clientes que tiene iFood por fuera del dataset.

Forma de recolección del conjunto de datos

El conjunto de datos contiene características sociodemográficas y firmográficas sobre 2240 clientes que fueron seleccionados al azar y contactados telefónicamente.

Infomación sobre el conjunto de datos

El conjunto de datos de iFood contiene en total 29 columnas, en las cuales se puede separar claramente en datos del cliente, los productos que comprar los clientes, las promociones que aceptaron o no los clientes, los diferentes tipos de compras y cantidad de visitas al sitio web en el último mes.

En los datos del cliente podemos observar datos como el ID, el año de nacimiento, nivel educativo, estado civil, ingreso familiar anual del cliente, cantidad de niños y adolescentes en el hogar, fecha de alta del cliente, número de días desde la última compra que realizó y si se quejó o no en los últimos dos años.

Year_Birth	Año de nacimiento del cliente
Education	Nivel de educación del cliente. Puede tomar los valores: Graduation, PhD, Master y 2n Cycle
Marital_Status	Estado civil del cliente. Puede tomar los valores: Single, Together, Married, Divorced, Widow, Absurd, Alone y YOLO
Income	Ingreso familiares anuales del cliente. Promedio del ingreso anual del cliente de los años 2018,2019 y parte del 2020.
Kidhome	Número de niños en el hogar del cliente. Puede tomar los valores: 0,1,2,3
Teenhome	Número de adolescentes en el hogar del cliente. Puede tomar los valores: 0,1,2,3
Recency	Números de días desde la última compra del cliente. Puede tomar los valores [0..99]

En los datos del producto que compran los clientes tenemos la cantidad de reales gastada en los diferentes productos que ofrecen las tiendas de iFood.

MntX	Cantidad de Reales gastadas en el producto X en los últimos 2 años. Donde X puede tomar el valor: Wines, Fruits, Meats, Fish, Sweet y Gold
------	---

En la sección de promociones tenemos datos sobre el cliente acepta o no una de las campañas propuesta por la empresa

AcceptedCmpX	Toma el valor 1 si el cliente aceptó la oferta en la campaña X. Donde X puede tomar el valor: 1,2,3,4,5
Response	Toma el valor 1 si el cliente aceptaría la próxima campaña (la sexta) y 0 en caso contrario

Por otra parte tenemos datos sobre el número de compras realizadas.

NumXPurchases	Número de compras realizada utilizando el medio X. Donde X puede ser: Deals, Web, Catalog y Store
---------------	--

Preguntas de interés

- ¿Que características tienen los clientes de la empresa?
- ¿Que características tienen los clientes que aceptan las campañas de la empresa?

Hipótesis

Para un supermercado es interesante saber qué tipos de usuarios son los que compran, para poder dirigir campañas publicitarias de productos específicos y así mejorar las ventas. Por ende, vamos a realizar un clustering con estos datos para conocer los diferentes perfiles de compradores y qué relaciones hay entre ellos.

Además se quiere saber si va a funcionar una sexta campaña publicitaria en todo el conjunto de usuarios, por lo que vamos a realizar diferentes modelos de clasificación/predicción basados en las respuestas telefónicas que dieron sobre si aceptarían o no la misma y así analizar la viabilidad de esta.

Preprocesamiento de datos

- Eliminamos los atributos *Z_Revenue* y *Z_CostContact* ya que ambas columnas tiene los mismos valores en todos los ejemplo, por lo que no son relevantes.
- Eliminamos los clientes con *Income* desconocido (22 clientes)
- Eliminamos los clientes que tenían el atributo *Income* menores a 14500 y los mayor a 88000 ya que tenían muy pocos elementos a partir de esos rangos.
- Se pasaron los valores de los atributos *Income*, y los atributos de dinero gastado (*MntWines*, *MntGoldProducts*, *MntFruits*, *MntFishProducts*, *MntWines*, *MntSweetProducts*) de Reales a Dólares para que sea mas fácil entender las magnitudes. El cambio de Real a Dólar se calculo como 4.252 Reales a 1 Dólar. Este valor viene calculado del promedio del valor del Dólar durante 2018 (3.653 Reales cada Dólar), 2019 (3.945 Reales cada Dólar) y 2020 (5.16 Reales cada Dólar). Usamos el promedio porque no sabemos con exactitud cuanto gano la persona cada año como para aplicar cada precio de Dólar para cada año.
- Creamos un nuevo atributo *Age* utilizando el atributo *Year_Birth* el cual indica la edad actual del cliente porque es mas fácil de entender la edad de una persona y hacer comparaciones que usando fechas.
- Eliminamos clientes que tenían mas de 90 años ya que eran muy pocos y ademas había clientes con mas de 120 años.
- Creamos un nuevo atributo *InRelationship* donde las personas con *Marital_Status* con valor "Together" o "Married" tienen como valor "1" y el resto de valores "0"
- Creamos un nuevo atributo *Total_Spent* el cual tiene la sumatoria de todos los productos que el cliente compró.
- Creamos un nuevo atributo *Children* a partir del atributo *Kidhome* y *Teenhome* para saber cuantos hijos hay en la casa y no tenerlo dividido en dos "tipos" de hijos cuando no hace tanta diferencia
- Creamos un nuevo atributo *Is_Parent* para saber si el cliente es padre o no. Se hizo usando el atributo *Children*. Para determinar rápidamente si es padre
- Agrupamos en 3 categorías el atributo *Education*. Ahora tiene la categoria de *Undergraduate* (Para educación "Basic" y "2n Cycle"), *Graduate* (Para "Graduation"), *Postgraduate* (Para "Master" y "Phd") para simplificar los valores del mismo.

Experimentos realizados

- Agrupamiento: Se comienza con un agrupamiento de los clientes para detectar que tipos de clientes tiene iFood y que características distintivas se pueden encontrar.
 - Se utiliza el algoritmo K-Medias para realizar el agrupamiento.
- Clasificación: Se construyen distintos modelos para poder predecir si los clientes aceptarán la última campaña (la sexta) y se hace una comparación entre los mismos.
 - Naive Bayes
 - Árbol de Decisión
 - Multi-perceptrón

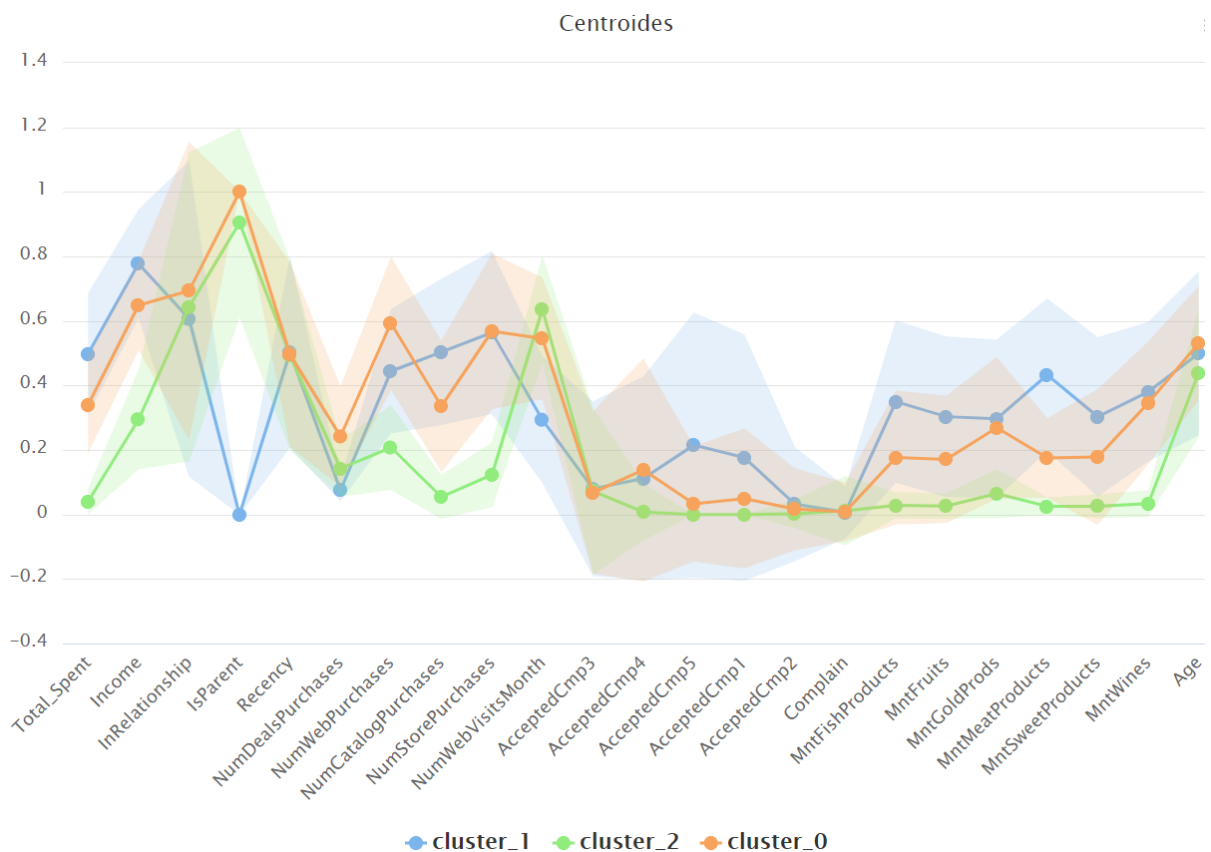
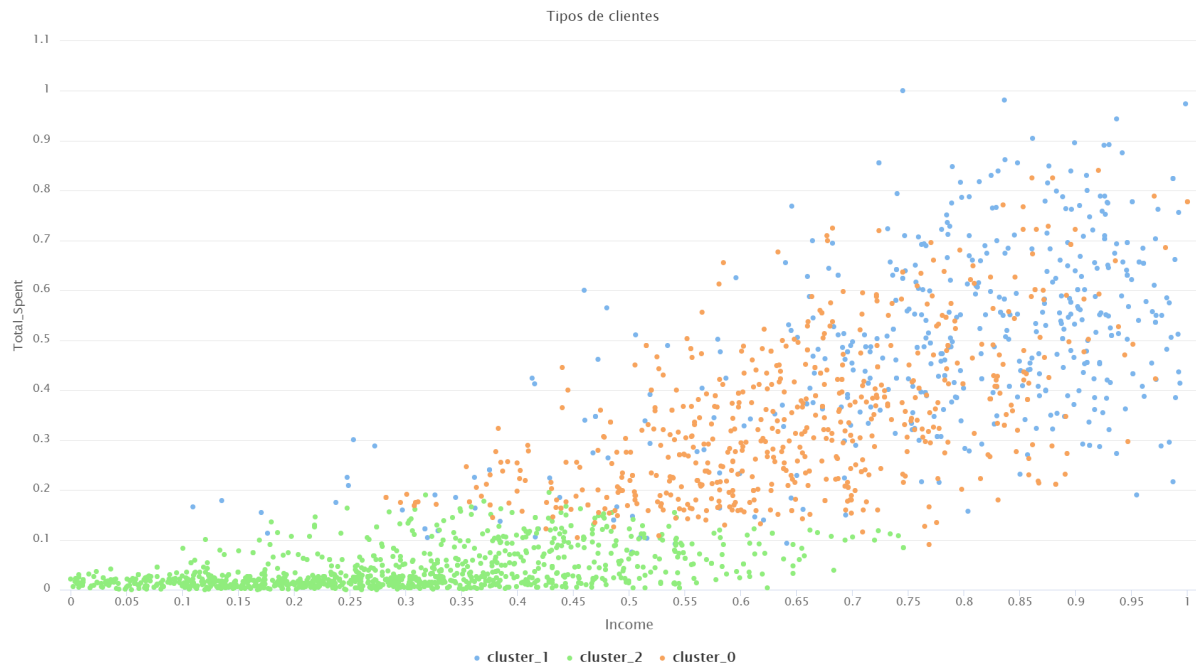
Agrupamiento

Se realizo un agrupamiento de clientes para saber que tipos de clientes tiene la empresa

Empezamos probando distintos valores de K para usar en K-medias y obtuvimos esta tabla:

Clustering.k	2	3	4	5	6	7	8
Davies Bouldin	-0.062	-0.077	-0.066	-0.059	-0.081	-0.072	-0.069

A pesar de que con un K = 6 obtendríamos el mejor clustering, preferimos utilizar un K = 3 (la segunda mejor opción) porque nos da una división mucho mas clara de los tipos de clientes con los que estamos tratando a cambio de una pérdida de precisión mínima en el clustering.



Análisis realizado

Rápidamente al ver el clustering se pudo ver que hay 3 tipos claros de clientes, aquellos con bajos (grupo verde, 1030 personas), medios (grupo naranja, 607 personas) y altos ingresos (grupo azul, 465 personas), por lo que nos centramos en encontrar que otras características distinguen a cada grupo.

- **Dinero Gastado:** La cantidad de dinero gastada por cada grupo tiene una correlación positiva con los ingresos que tienen.
- **Edad:** En promedio el grupo 1 es el mas joven seguido por el grupo 3, esto nos dice que el grupo 2 son clientes mayores y de ingresos medios.
- **Compras totales:** En este caso el grupo 2 es el que mas compras ha realizado, seguido por el grupo 3. Lo que nos dice es que aunque el grupo 2 no es el que mas dinero ha gastado si es el que mas compras realiza, y se los podría considerar los clientes regulares de la empresa.
- **Campañas aceptadas:** En general se puede ver que el grupo que tiene una mayor respuesta positiva con las campañas es el grupo 3, a su vez el grupo 2 suele tener una aceptación de campañas menor, y el grupo 1 es el que tiene una respuesta mayormente negativa hacia las campañas
- **Familia:** En cuanto a estar en pareja o no, todos los grupos tienen un balance entre personas solteras y en pareja. Cuando hablamos de hijos, los grupos 1 y 2 no tienen nada que destaque, pero el grupo 3 destaca por ser un grupo donde hay pocas personas con hijos.
- **Tipos de compras:** El grupo 1 es un grupo que hace pocas compras de todo tipo, dando la idea de que son un grupo de clientes que no suelen comprar seguido en iFood, en el caso del grupo 2 suelen hacer muchas compras con ofertas, a través de la web y también en las tiendas, son los clientes regulares como se menciono anteriormente. En el caso del grupo 3 vemos que no son de comprar con ofertas debido a sus ingresos y que se destacan por ser un grupo que compra mucho por catalogo, ademas de también comprar en las tiendas.

Resumen

Estos son los grupos encontrados:

- Grupo 1: Personas en general mas jóvenes que otros tipos de clientes, no son clientes habituales de iFood y se centran en visitar la web para compras ocasionales, tienen bajos ingresos y tiene pocas compras en iFood y por lo tanto poco dinero se obtiene de estas personas.
- Grupo 2: Los clientes regulares de iFood, son gente adulta de familia con ingresos medios y medios-altos que compran en iFood en persona y online, son de aprovechar las ofertas que encuentran y son el grupo que mas compras ha realizado.
- Grupo 3: Los clientes que mayor ganancia le dan a iFood, no son tan regulares como los del grupo 2, pero igualmente son un grupo muy activo para la empresa. Son personas adultas y la mayoría no tienen hijos. No son de comprar utilizando ofertas, compran mucho por catalogo y son el grupo que mas suele aceptar las campañas de iFood.

Conclusión

Si iFood busca aumentar la cantidad de personas que aceptan campañas debe empezar por atraer al grupo 2 y mantener o aumentar la atracción del grupo 3. El grupo 1 no vale la pena lo suficiente como para intentar orientar una nueva campaña hacia ellos.

Clasificación

Utilizaremos el atributo *response* el cual esta basado en respuestas telefónicas que dieron los clientes sobre si aceptarían o no la próxima campaña (la sexta) para saber si va a tener una buena recepción.

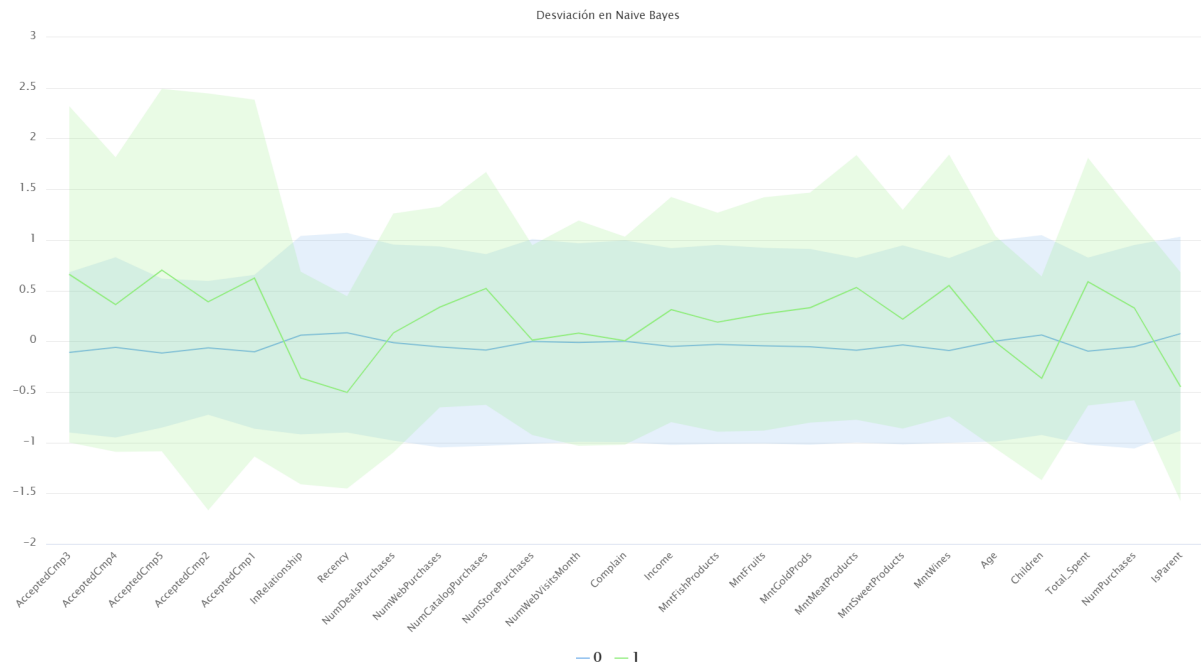
Los siguientes modelos fueron contruidos utilizando k-fold cross validation con k=10 para tener una estimación mas precisa del rendimiento de los modelos.

Modelos contruidos

Naive Bayes

Se construyó un modelo de Naive Bayes para clasificar clientes y saber si van a aceptar o no la campaña. Se normalizaron los atributos antes de construirlo.

Las desviaciones en el modelo nos quedaron de la siguiente manera:



La matriz de confusión del modelo es la siguiente:

accuracy: 80.54% +/- 2.42% (micro average: 80.54%)

	true 1	true 0	class precision
pred. 1	154	263	36.93%
pred. 0	146	1539	91.34%
class recall	51.33%	85.41%	

Hay dos clases, la primera (Clase 1) es la clase de las personas que si aceptaron la campaña 6, la segunda (Clase 0) son los que la rechazaron.

Analizando el modelo podemos notar que tanto la precisión como el recall de la clase 1 dejan mucho que desear.

Atribuimos este resultado a que en el dataset tenemos aproximadamente un 75% de clientes de clase 0, y solo un 25% de clase 1, por lo que es mas difícil de determinar de forma general que es lo que estas personas tienen en común y por lo tanto armar un modelo que los abarque a todos. En el caso de la clase 0 es mas fácil, ya que hay mas muestras con las que ajustar el modelo.

Consideramos que la utilidad de este modelo puede darse para determinar con bastante precisión cuando alguien NO aceptara la campaña 6, igualmente la utilidad esta en duda ya que los clientes de importancia son los de clase 1, que podrían ser objetivo de algún marketing específico.

Analizando las desviaciones podemos ver que las personas que aceptan las campañas tienden a ser personas que ya habían aceptado campañas y ademas que en promedio están gastando mas dinero en iFood, ademas de que en promedio tienen menos hijos. Esto tiene una similitud con el grupo de gente adinerada descubierto en el clustering.

Árbol de Decisión

Se construyo un árbol de decisión con el dataset para realizar una comparación de accuracy con Naive Bayes y el próximo modelo.

El árbol resultante es el siguiente:



Y la matriz de confusión que obtuvimos es la siguiente:

accuracy: 87.16% +/- 2.85% (micro average: 87.16%)

	true 1	true 0	class precision
pred. 1	72	42	63.16%
pred. 0	228	1760	88.53%
class recall	24.00%	97.67%	

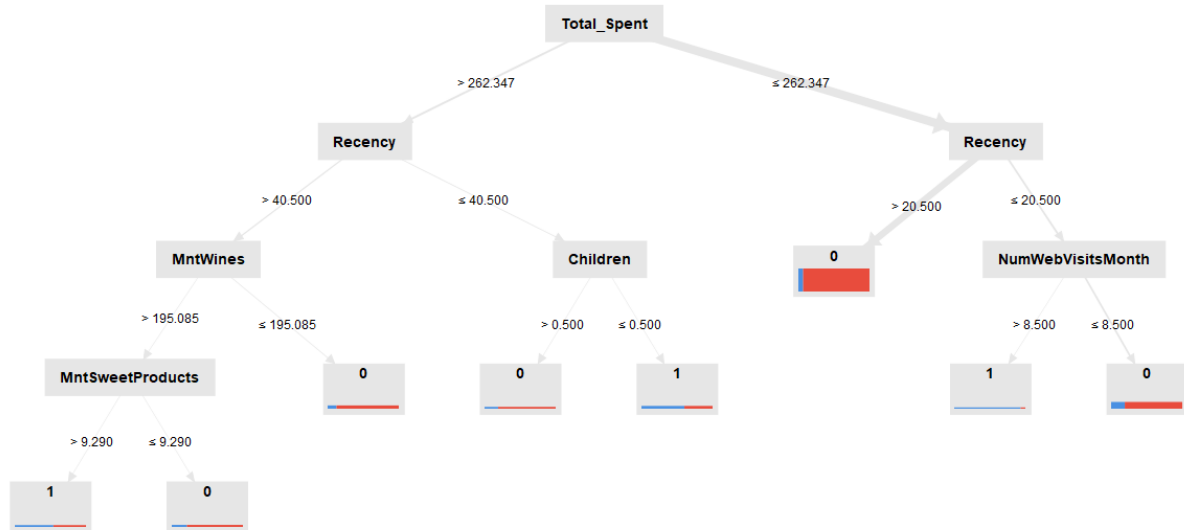
En este caso el modelo tiene una mejor precisión de la clase 1 que Bayes pero un peor recall para la misma.

A pesar de que el accuracy del modelo es mayor, no se puede decir que el modelo es mejor que el de Bayes ya que clasifica erróneamente a la mayoría de clientes de clase 1 que son los importantes para esta clasificación, ya que podrían ser el objetivo del marketing a futuro para la campaña.

Analizando los atributos que construyen el árbol se puede ver que los clientes que son regulares además de tener muchas visitas online son los que aceptan la nueva campaña, aun así los dos atributos más importantes son los que dicen si una persona aceptó la última campaña (la 5) y la 3era. En el caso de que esas dos campañas no hayan sido aceptadas, hay pocas posibilidades de que la persona acepte la 6ta campaña. Esto hace difícil de usar el árbol para clasificar nuevos clientes que no hayan tenido contacto con las campañas anteriores, por lo que se procederá a construir otro árbol de decisión pero sin los atributos de las 5 campañas hechas hasta ahora, para determinar si se puede hacer un modelo similar que no necesite de las campañas anteriores para predecir.

Árbol de decisión sin campañas anteriores

Se construyó un árbol de decisión quitando los atributos de aceptación de las 5 campañas anteriores y quedó de la siguiente manera:



La matriz de confusión resultante es la siguiente:

accuracy: 86.44% +/- 1.52% (micro average: 86.44%)

	true 1	true 0	class precision
pred. 1	59	44	57.28%
pred. 0	241	1758	87.94%
class recall	19.67%	97.56%	

Se puede ver que es un modelo peor que el mismo con los atributos de las campañas anteriores.

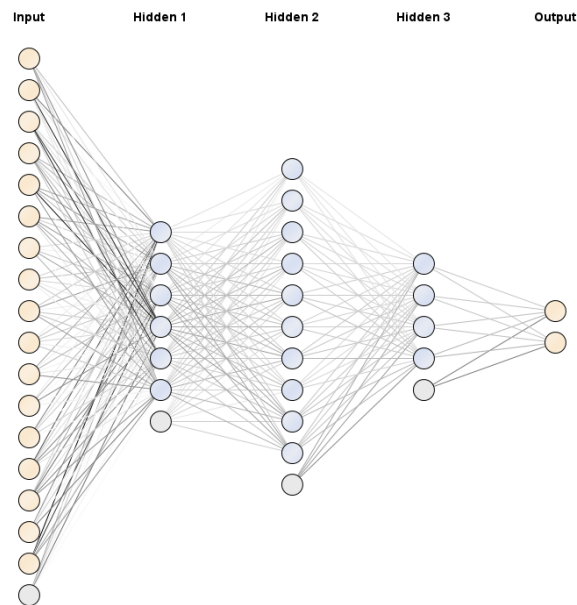
Analizando los atributos del árbol se ve que el atributo mas importante es **Total_Spent** que dice que si una persona gasta menos de cierta cantidad (bastante alta porque tiene un valor de 0.966 y los valores están normalizados) es bastante seguro de que no aceptara la campaña, esto se condice con el clustering y con el modelo Bayes donde vemos que las personas que mas ingresos tienen son las que mas campañas aceptan. El modelo ademas dice que de esas personas que mas dinero gastan en iFood y aceptan campañas tienen pocos hijos y son de comprar vino y dulces. Esto también coincide con la gente adinerada vista anteriormente.

En el caso de que se quiera clasificar a nuevos clientes no sería útil ninguno de los árboles construidos, el anterior por lo mencionado previamente sobre la no-interacción de los nuevos clientes con las campañas previas. Y este nuevo árbol porque el atributo mas importante es cuanto ha gastado una persona en total en los últimos 2 años, que es imposible de tener para un cliente nuevo.

Red Neuronal

Se construyo un multiperceptrón para clasificar los clientes y poder hacer una comparación con los modelos previos.

La estructura de la red es la siguiente:



La estructura se decidió luego de probar distintos valores manualmente y encontrando que la misma daba una precisión mejor que en otros casos.

El atributo **Education** que es polinomial fue dividido en 3 atributos binarios: **Education=Undergraduate**, **Education=Graduate**, **Education=Postgraduate**

Una vez construida y entrenada la red, la matriz de confusión resultante es la siguiente:

accuracy: 88.15% +/- 1.20% (micro average: 88.15%)

	true 1	true 0	class precision
pred. 1	144	93	60.76%
pred. 0	156	1709	91.64%
class recall	48.00%	94.84%	

Este modelo es un intermedio si se lo compara con los demás, por un lado tiene un mal recall para la clase 1 pero es mejor que ambos árboles construidos, por otro lado tiene una precisión de clase 1 que es bastante mejor que el modelo de Bayes y mejor que uno de los árboles. Si se habla de la precisión y recall de la clase 0 podría considerarse el modelo mas balanceado y mas útil para determinar que personas NO van a aceptar la próxima campaña.

Debido al balance que presenta en precisión y recall (ademas de su accuracy general que es bastante buena) se considera que este modelo es el que podría utilizarse para clasificar el resto de clientes que posee iFood.

Comparación entre modelos

Se armo una tabla comparativa para los modelos, se calculo el F-Score de cada clase con la siguiente formula: $F_1 =$

$$2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Modelo	Accuracy	Precisión Clase 0	Recall Clase 0	F-Score Clase 0	Precisión Clase 1	Recall Clase 1	F-Score
Naive Bayes	80.54%	91.34%	85.41%	0.8827	36.93%	51.33%	0.4295
Árbol con campañas	87.16%	88.53%	97.67%	0.9287	63.16%	24%	0.3478
Árbol sin campañas	86.44%	87.94%	97.56%	0.925	57.28%	19.67%	0.2928
Red Neuronal	88.15%	91.64%	94.84%	0.9321	60.76%	48%	0.5363

Conclusión

Si utilizáramos alguno de estos modelos de clasificación para clasificar los clientes en iFood se haría claramente con el modelo de multi-perceptrón ya que es el que mejor clasifica a los clientes de clase 0 y los F-score de ambas clases son los mejores comparados contra todos los demás modelos. A pesar de esto la clasificación tiene mucho para mejorar y sería muy útil tener información sobre más clientes para poder entrenar los modelos mejor.

Además podemos ver que la clasificación que se puede realizar sobre los clientes tiene muchos datos que necesitan que cada individuo tenga un tiempo suficiente como cliente de iFood antes de poder utilizar sus datos para clasificarlo. Por lo que toda clasificación posible con este dataset servirá para el resto de clientes que tenga iFood y no para los nuevos.

Conclusión Final


iFood tiene grupos bastante diferenciados de clientes que utilizan sus servicios, son 3 grupos bien claros de clientes. El primer grupo son personas que interactúan poco con iFood, bajos ingresos y no aceptan casi ninguna campaña, el segundo grupo son los clientes regulares de la empresa, aceptan algunas campañas y son los que mas compras realizan en iFood, el tercer grupo son personas de alto poder adquisitivo que son los que mas campañas aceptan en iFood y que gastan mucho dinero también.

Con la clasificación se pudo ver que las personas que aceptan las campañas son aquellas que aceptaron previamente otra campaña (en especial la 5ta y la 3era) y ademas aquellos que previamente hayan gastado mucho dinero.

La conclusión que se obtiene es que el marketing se puede hacer especifico para cada grupo de clientes con los siguientes objetivos:

- Empezar a captar gente del grupo de pocos ingresos ya que son muchos y es donde mayor cantidad de clientes regulares se pueden obtener
- Aumentar la cantidad de clientes regulares que aceptan campañas
- Mantener o aumentar la cantidad de clientes de altos ingresos que gastan dinero en iFood y que aceptan casi todas las campañas

Referencias

 <https://github.com/nailson/ifood-data-business-analyst-test/blob/master/iFood%20Data%20Analyst%20Case.pdf>

Customer Personality Analysis

Analysis of company's ideal customers

 <https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis>

