Segmentación de clientes y predicción: Supermarket Sales Birmania 2019

MASTER EN BIG DATA Y BUSINESS ANALYTICS
2019-2020

24 SEPTIEMBRE

JUAN MANUEL ORTIZ RAMÍREZ

Indice

1. PRESENTACIÓN

Permitanme introducirme, soy Juan Manuel Ortiz, con 34 años en las espaldas y madrileño de cuna. Actualmente soy Team Leader de un magnifico equipo de 10 personas, en el European Repair Centre del grupo Glory Global en Madrid. Mi personalidad activa y dinámica y mis insaciables ganas de aprender y mejorar, hicieron que me interesase por dar un giro a mi carrera profesional y educativa, cayendo en el atractivo mundo del Big Data.

Mi abanico de conocimientos, mezclando campos técnicos, sociales y de gestión, me hacen tener una visión amplia y diferente en cualquier aspecto de la vida, que intento aplicar en todos los ambitos.

"La actitud es una pequeña cosa, que hace una gran diferencia" Winston Churchill

2. INTRODUCCIÓN

Mediante el estudio y analisis del conjunto de datos escogido, podré aplicar y poner en práctica los distintos conocimientos aprendidos a lo largo del temario del curso.

Ello conllevará enfrentarme a nuevos retos y dificultades que no han surgido en prácticas anteriores, pero que a su vez implicará seguir aprendiendo y aplicando nuevos conocimientos, técnicas o metodos resultantes del autoaprendizaje.

El hecho de escoger un conjunto de datos relacionado con los datos de un supermercado, se debe principalmente, a que es un caso muy práctico, que facilmente puede darse en la realidad y puede ser aplicado a multiples sectores.

El código a utilizar es integramente en lenguaje R, ya que su facilidad de uso y diversidad de funciones, hacen que sea completamente adecuado para este trabajo. Otras herramientas como Excel, Word tambien serán necesarias para el correcto desarrollo.

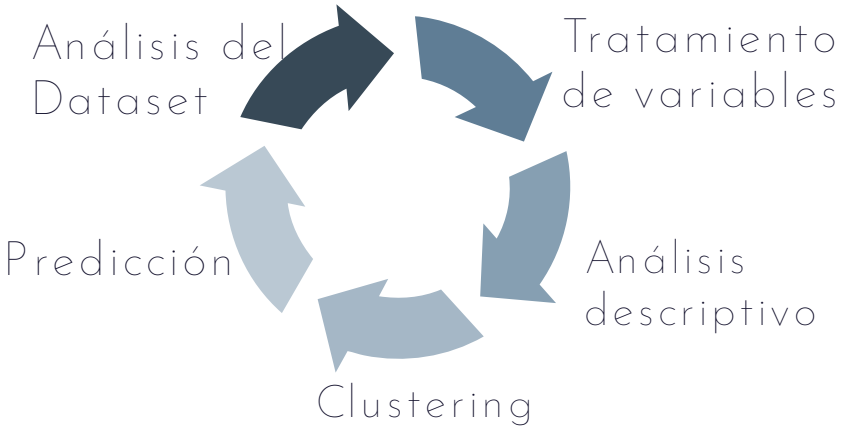
3. METODOLOGÍA Y OBJETIVOS

El fín de este TFM es realizar un analisis sobre el conjunto de datos obtenidos de un supermercado de Birmania, que nos permita realizar una segmentación de los clientes. Para ello se escogerá de entre distintos modelos el que creamos que es más conveniente para nuestro fín.

El dataset trata los datos de supermercados de tres diferentes ciudades, a falta de más detalle por el creador, voy a considerar que son de la misma compañía. Podría considerar que son distintas compañias y esto haría que el informe tuviese un enfoque distinto.

Una vez tengamos los distintos grupos bien diferenciados, procederé a realizar un analisis predictivo que nos permita hallar con la máxima precisión posible cada uno de los grupos.

El metodo de trabajo empleado se puede visualizar de manera rapida en el esquema inferior. Para ello primero trataremos de realizar un analisis del conjunto de datos. El conocer las caracteristicas del conjunto, nos permitirá tomar decisiones sobre el tratamiento de las variables, con un analisis descriptivo profundo podremos hallar y ampliar la información que esconden los datos. Con todo lo anterior, podremos proceder a realizar la segmentación de clientes, y posteriormente la predicción de dichos resultados.



"Este método de trabajo, no es rígido, a medida que se avanza y se obtiene más información, es necesario retroceder algún paso para realizar cambios o mejorar el proceso."

Este informe no ahonda en los detalles en cuanto a código, ha sido realizado íntegramente en lenguaje R, con la necesidad de uso de múltiples librerías y recursos, que se pueden ver más detalladamente en los anexos.

4. ANALISIS DATASET ORIGINAL

El conjunto de datos utilizado para este trabajo ha sido obtenido del siguiente enlace:
https://www.kaggle.com/aungpyaeap/supermarket-sales

Fue subido a la plataforma KAGGLE por el usuario Aung Pyae el 2019-05-27
Este usuario describe el dataset de la siguiente manera:

"El crecimiento de los supermercados en la mayoría de las ciudades pobladas está aumentando y las competencias de mercado también son altas. El conjunto de datos es una de las ventas históricas de la empresa de supermercados que ha registrado datos en 3 sucursales diferentes durante 3 meses. Los métodos de análisis de datos predictivos son fáciles de aplicar con este conjunto de datos."

Downloads: 21,2k | Views: 104k | Kernels: 26 | Discussions: 7

A continuación, se recoge una tabla con un resumen del conjunto de datos:

1000 valores | 17 variables | 9 Character | 7 Numeric | 1 Integer

Variable	Tipo	Descripción
Invoice.ID	Chr	Nº de identificación de factura de comprobante de venta generado automáticamente
Branch	Chr	Sucursal del supermercado (hay 3 sucursales A, B y C)
City	Chr	Localización de los supermercados. Yangon, Naypyitaw y Mandalay
Customer.Type	Chr	Tipo de clientes. Member clientes que usan tarjeta de socio y Normal
Gender	Chr	Género de los clientes. Female/Male
Product.line	Chr	Grupos de categorías generales de los ítems. 5 categorías
Date	Chr	Fecha de compra (enero 2019 – marzo 2019)
Time	Chr	Hora de compra (10:00-21:00)
Payment	Chr	Forma de pago usado. Efectivo, Tarjeta, Pago Digital
Unit.Price	Num	Precio de cada producto en $
Rating	Num	Calificación de la experiencia de compra del cliente. 0-10
Tax.5.	Num	Impuesto del 5% por cada compra
Total	Num	Ingreso total
Cogs	Num	Coste de Bienes Vendidos
Gross.margin.percentage	Num	Porcentaje del margen de ingresos
Gross.income	Num	Ingresos Brutos
Quantity	Int	Número de productos comprados por clientes

Realizando un breve análisis EDA, podríamos adivinar que, por la singularidad de los datos, sus proporciones y errores encontrados, este podría tratarse de un dataset artificial, creado sólo como un ejemplo para practicar y realizar ejercicios. En la realidad, sería difícil encontrar datos de supermercados con estas proporcionalidades tan exactas.

En el anexo podremos comprobar con más detalle la frecuencia de los distintos niveles de las variables categóricas, la correlación entras las variables, los niveles más comunes de las variables categóricas, el tamaño que ocupa en la memoria cada variable e histogramas de las variables numéricas.

Valores equitativamente distribuidos

Branch	34%	33,2%	32,8%
City	34%	33,2%	32,8%
Customer type	50,1%	49,9%	
Gender	50,1%	49,9%	
Product Line	17,8%	17,4%	17%
	16,6%	16%	15,2%
Quantity	De 1-10 con 8,5% a 11,9%		
Payment	34,5%	34,4%	31,1%

Gross.margin.percentage	4,76% Valor constante
Cantidad de NA's	Ninguno

City = Branch

Tax.5. = Gross.Income

- Haciendo un simple análisis de la estructura de los datos, hallamos que las variables están muy balanceadas, y sus datos equitativamente distribuidos, entre sus distintas categorías/valores.
- Esto puede ser una dificultad añadida a la hora de encontrar patrones cuando vayamos a realizar la segmentación de clientes.
- "Gross.margin.percentage": Es una variable constante que no aporta información.
- El dataset no contiene ningún valor ausente.
- "City" y "Branch" son similares, pero City puede aportar algo más de información extra.
- "Gros.Income" parece una variable errónea.
- La variable "Quantity" también se encuentra distribuida equitativamente entre sus diversas posibilidades, ya que los porcentajes que tienen se encuentran en un rango de entre 8.5% y 11.9%

5. Tratamiento de variables

Para poder manejar mejor el dataset, y obtener información más precisa/útil, es necesario proceder a la transformación, eliminación o generación de variables. Este es un primer tratamiento de los datos, puesto que posteriormente para los diferentes procesos que apliquemos, deberemos realizar otras transformaciones.

- Product.Line (Character)

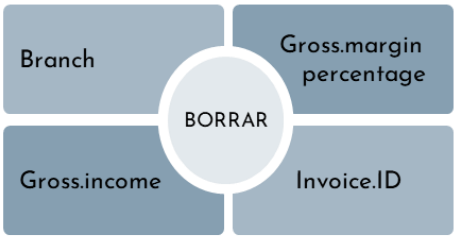
Modificamos las 5 categorías de línea de producto para hacerlas más simples, sencillas y facilitar su comprensión a primera vista

- Date (Character)
- Time (Character)

» Datetime(Posixct)

» Day (numeric)
» Month (numeric)
» Week (numeric)
» Hour (numeric)
» Daynum (numeric)

Uno la variable Date y Time, y de la resultante, extraemos otras variables interesantes que pueden sernos útiles en el futuro, como el día de la semana, mes, semana, hora y día del mes en el que se realizaron las compras.



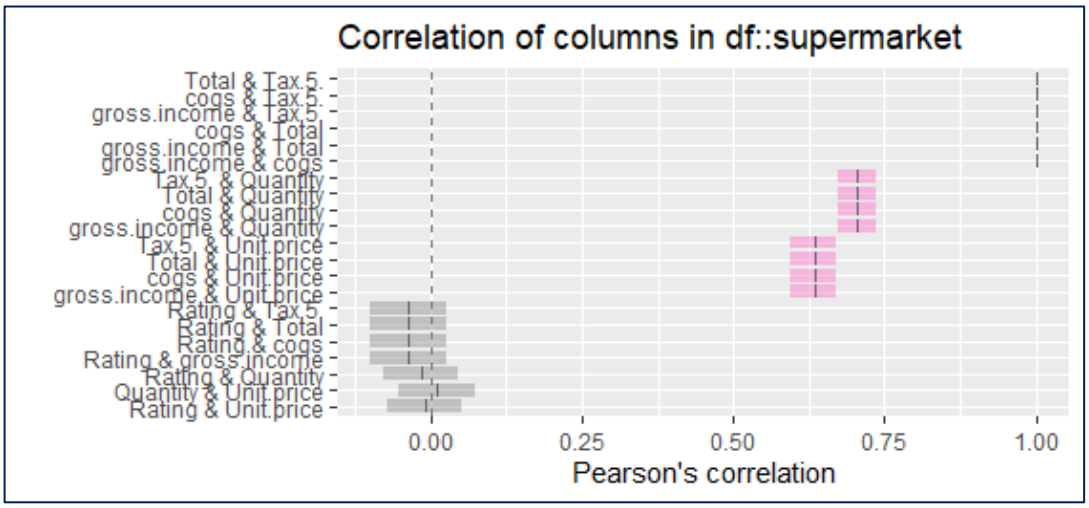
Cómo comenté anteriormente, algunas variables carecen de sentido y por ese motivo son eliminadas:

- Branch es eliminada por ser similar a City
- Gross.margin.percentage es constante y no aporta valor
- Invoice.ID no aporta valor al ser un valor meramente identificativo de la compra
- Gross.income lo consideramos como una variable erróneamente creada al ser el mismo valor que Tax.5

- Temp (numeric)

Generamos una nueva variable que contiene la temperatura media de cada día, en cada una de las tres ciudades estudiadas (Yangon, Mandalay y Naypyitaw) .
Estos valores han sido obtenidos previa solicitud al NOA y con un Token mediante la web https://www.ncdc.noaa.gov/cdo-web/ . Con los valores Max y Min de cada día generé la temperatura media. Algunos valores "missing" de los facilitados por NOA fueron completados manualmente, extrayendo la información de https://sy.freemeteo.com/.

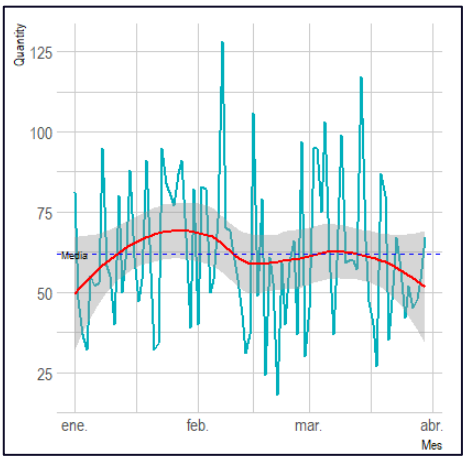
* Se adjunta Script temperatura.R donde se detalla el proceso de unificación y tratamiento de los datos obtenidos



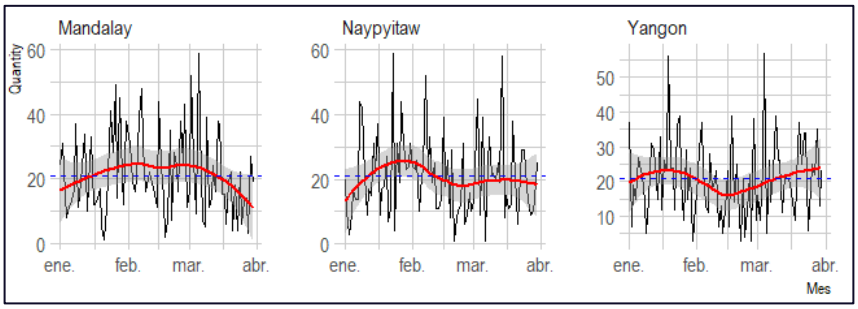
6. Análisis descriptivo

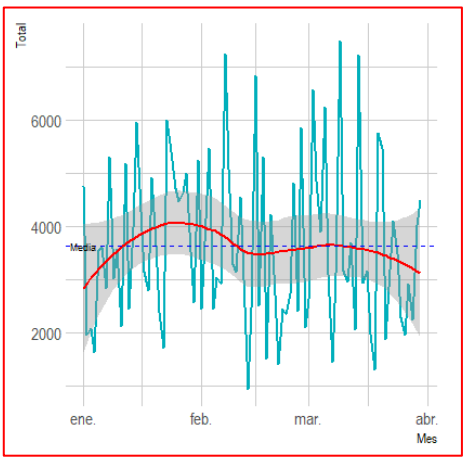
5.1 ANALISIS DESCRIPTIVO EN FUNCION DEL TIEMPO

Como anticipé anteriormente, en los cuatro gráficos inferiores, podemos comprobar como los valores de las variables están distribuidos de manera muy equitativa. Los valores medios, tanto de las cantidades vendidas (Quantity), como del total de ingresos (Total) en cada ciudad, fueron muy similares a lo largo de los tres meses. La mayoría de los valores se encuentran entorno a la media, con escasos valores atípicos.

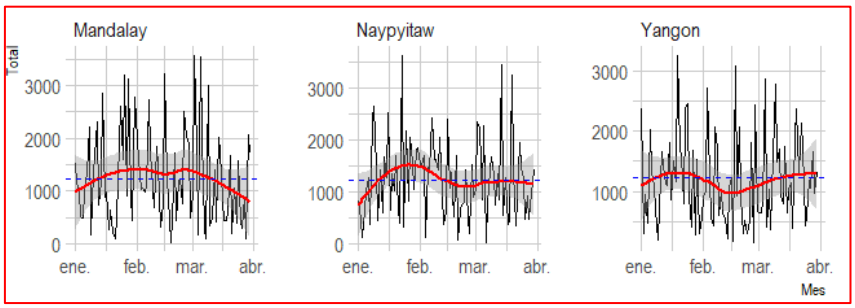


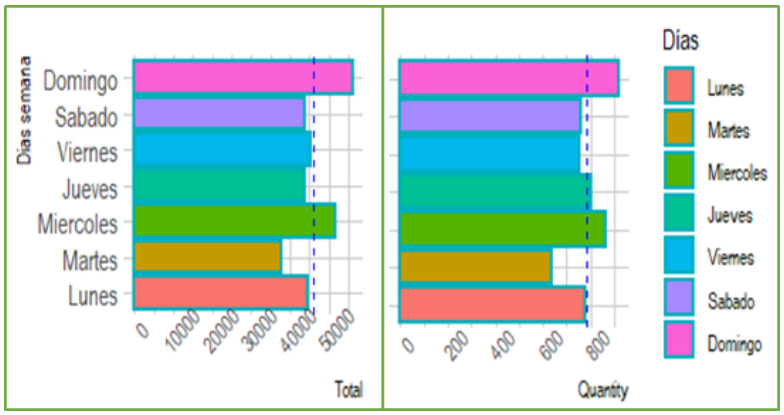
Cantidades Totales





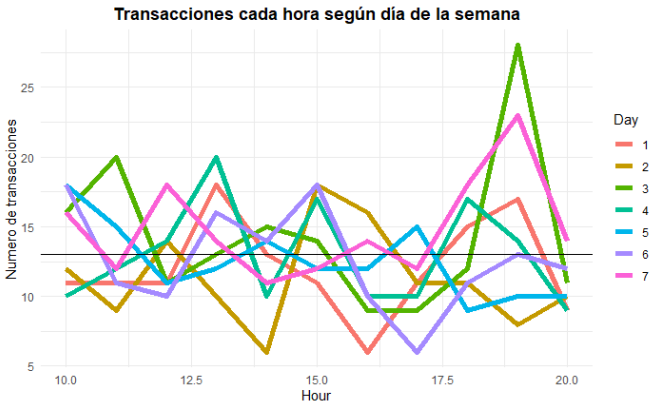
Ingresos totales





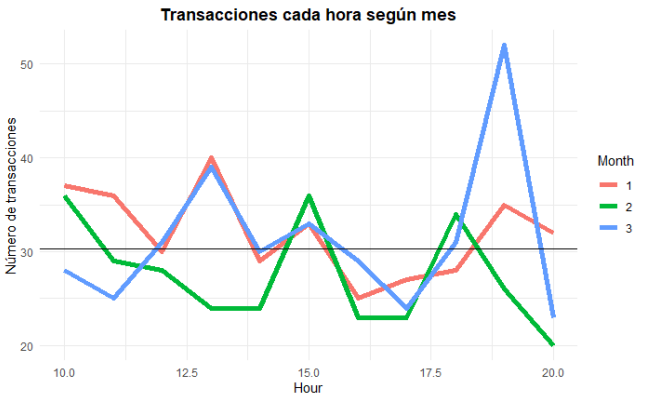
En función de los días de la semana, observamos:

- El miércoles y domingo destacan como los días con más ingresos y más unidades vendidas
- El martes es el día con peores resultados.
- El resto de días, se encuentran entorno a la media.



Observando el gráfico de transacciones cada hora según el día de la semana, apreciamos:

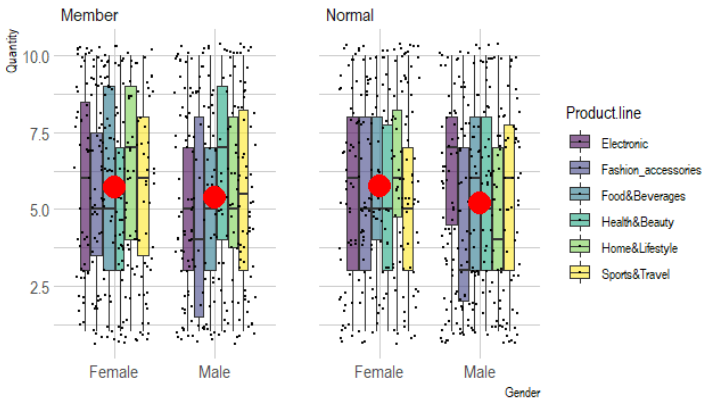
- El domingo y miércoles a las 19:00 se dan los picos máximos de compras.
- Por el contrario, el martes a las 14:00, el lunes a las 16:00 y el sábado a las 17:00, son los momentos con menos compras de la semana



Transacciones a la hora según el mes:

- Durante el mes de enero el número de transacciones se sitúan entorno a la media, con una tendencia decreciente.
- Febrero resulta el mes con peores resultados de ventas, situándose por debajo de la media.
- Marzo registra un gran volumen de ventas a las 19:00.

5.2 ANALISIS CLIENTE Y PRODUCTO

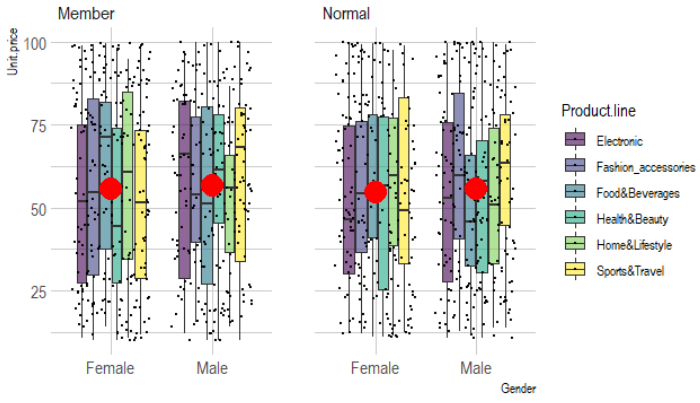


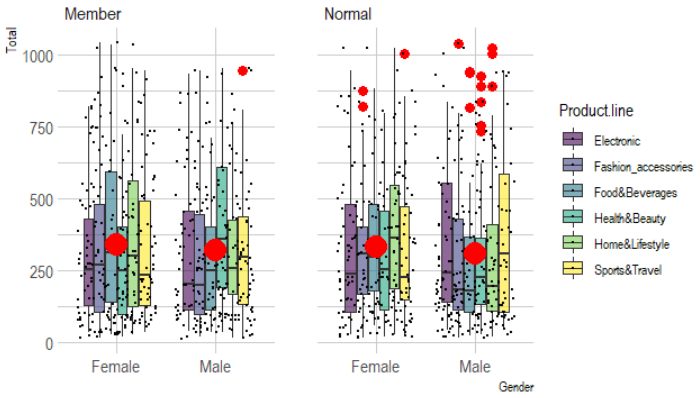
Analizando las cantidades de cada venta según el producto y tipo de cliente:

- Tanto por género, como por tipo de consumidor, las medias de las unidades vendidas en cada transacción son similares, siendo las mujeres quienes compran ligeramente más unidades..
- Por producto, parecen comprarse más unidades de productos tipo "Fashion" y "Home" por parte de los clientes "Miembros"

Analizando el precio unitario de venta según el producto y tipo de cliente:

- Las medias se mantienen similares.
- En clientes Miembros los productos de tipo "Fashion", "Food" y "Home" tienden a alcanzar precios más elevados, y en clientes Normales, destacan los productos tipo "Sport" entre las mujeres y los de tipo "Fashion" entre los hombres con precios mayores.



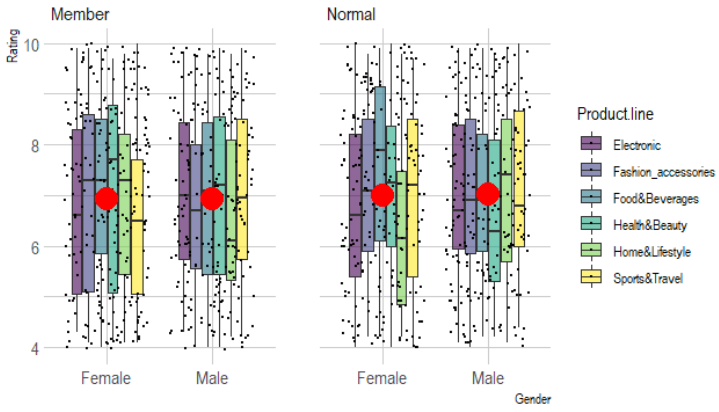


Analizando el Total de ingresos según el producto y tipo de cliente:

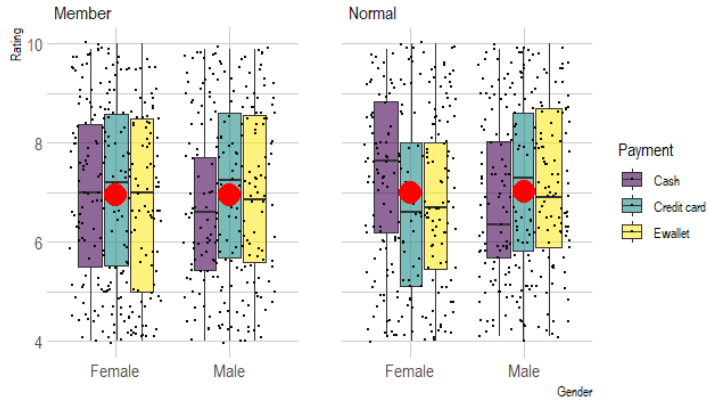
- Apreciamos posibles outliers en los consumidores Normales en productos de tipo "Food" y "Health" y "Home", donde menos ingresos se obtienen.
- En los consumidores Miembros y mujeres destacan con más ingresos los productos tipo "Food" y "Home", mientras que en los hombres destacan los de tipo "Health"

Analizando las valoraciones de cada venta según el producto y tipo de cliente:

- Destaca por encima de todos los productos tipo "Food" con mejores valoraciones,
- Opuestamente el producto peor valorado es de tipo "Home"
- En ambos casos son clientes Normales y Mujeres.



6.3 ANALISIS CLIENTE Y FORMA DE PAGO

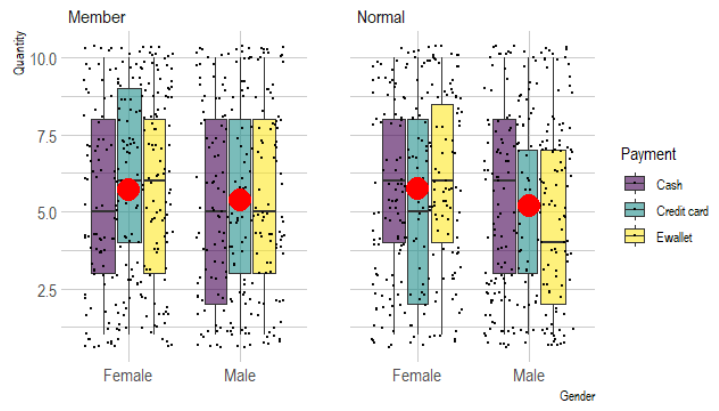


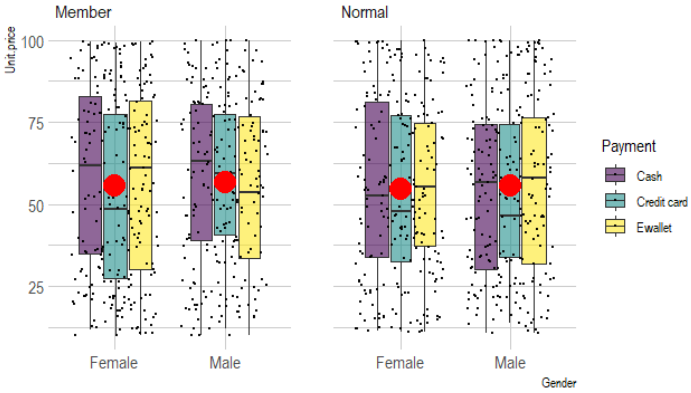
Analizando la valoración de las compras según el cliente y forma de pago:

- La media se mantiene similar, independientemente de cada cliente o medio de pago
- Se obtienen valoraciones algo más elevadas entre los clientes "Miembros", a excepción de los pagos realizados en "Cash" por mujeres entre los clientes "Normales".

Analizando las cantidades de cada venta según cliente y forma de pago:

- Las mayores unidades se dan entre las compras realizadas por las mujeres tanto "Miembros" como "Normales"
- Destacan con más unidades realizadas por compra las que son pagadas mediante Tarjeta de Crédito, por Mujeres de tipo Miembro.



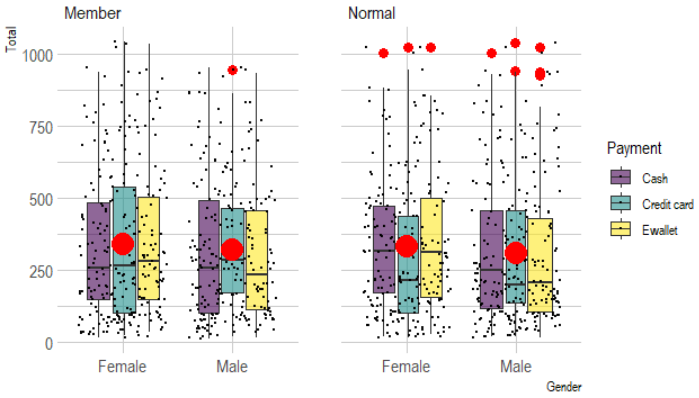


Analizando el precio unitario de venta según cliente y forma de pago:

- Las medias son similares, manteniendo la tónica de los datos.
- Destacando precios más elevados en las mujeres, y con pagos realizados mediante "Cash", tanto si son de tipo "Miembro" o "Normal"

Analizando el Total de Ingresos obtenidos según cliente y forma de pago:

- Aunque las medias son similares, se observan más compras que generan menos ingresos en lo hombres de tipo "Normal", esto provoca algunos outliers debido a que algunas pocas de las compras generan ingresos muy grandes.

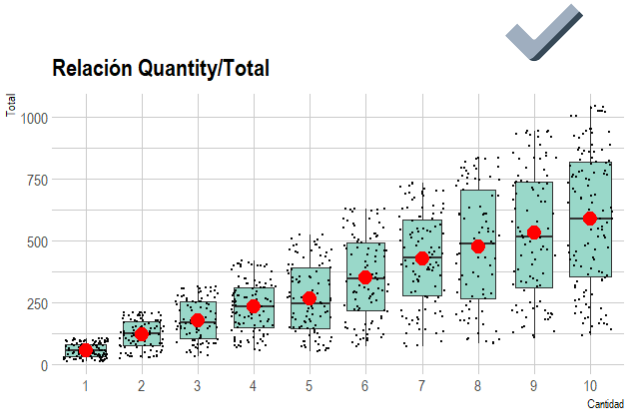


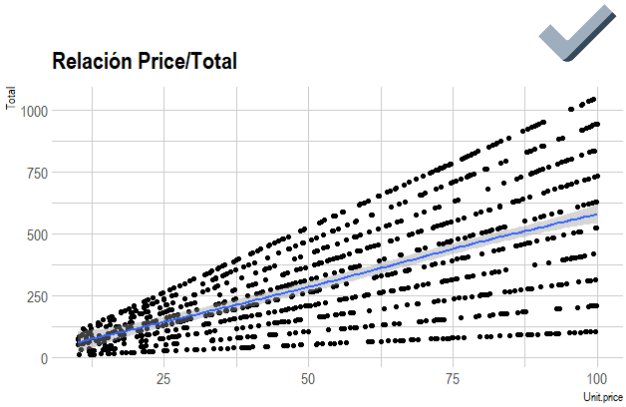
"Una vez analizados estos gráficos descriptivos tanto de los clientes, productos y formas de pago, es inevitable, no volver a llegar a la conclusión de que el conjunto de datos, no es especialmente realista, se aprecia que las medias se mantienen en los mismos valores, y los datos bien distribuidos, de manera equivalente. Así, a primera vista, es difícil llegar a conclusiones evidentes sobre los clientes y productos y llegar a relaciones que se puedan generar entre las distintas variables"

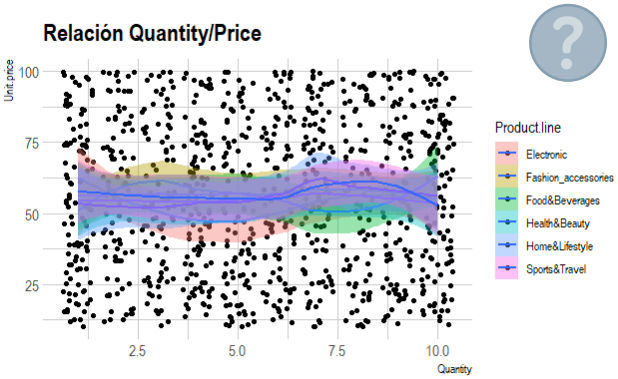
6.4 RELACIÓN ENTRE VARIABLES

Cómo ya se apuntaba en el gráfico de correlaciones obtenido del análisis del Dataset, hay dos relaciones directamente proporcionales entre las siguientes variables:

- Entre "Quantity" y "Total", se aprecia que a medida que aumenta la cantidad de productos vendidos, los ingresos aumentan.
- Entre "Price" y "Total", según se incrementa el precio, el nivel de ingresos también aumentan.







Se ha intentado comprobar si hay algún tipo más de relación, que pudiera ser obvia, como se podría suponer entre la cantidad vendida y el precio, pero como se observa en el gráfico, no hay ningún tipo de relación.

Cuando introducimos el componente del tipo de producto, se aprecian ya algunas diferencias:

- Productos de moda: menos unidades vendidas, pero a un mayor precio
- Productos de electrónica, salud/belleza y alimentación: se venden más unidades a precios más moderados.

7. CLUSTERING

Cómo uno de los objetivos de este TFM, voy a aplicar las diversas técnicas de clustering para comprobar si existe algún tipo de segmentación de clientes en nuestros datos.
Para ello voy a optar por probar con dos modelos que utilicen el método no jerárquico, como son:

- K-Means: es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano.
- PAM: es un método de clustering muy similar a K-means, ambos agrupan las observaciones en K clústeres, donde K es un valor preestablecido. En K-medoids, cada clúster está representado por una observación presente en el clúster (medoid).

Y otro modelo con el método jerárquico:

- Hierarchical Clúster Analysis: es un método de agrupamiento que se basa en encontrar jerarquías en los datos de entrada a partir de generar grupos basados en la cercanía o semejanza de tales datos. El aglomerativo es un acercamiento ascendente: cada observación comienza en su propio grupo, y los pares de grupos son mezclados mientras uno sube en la jerarquía

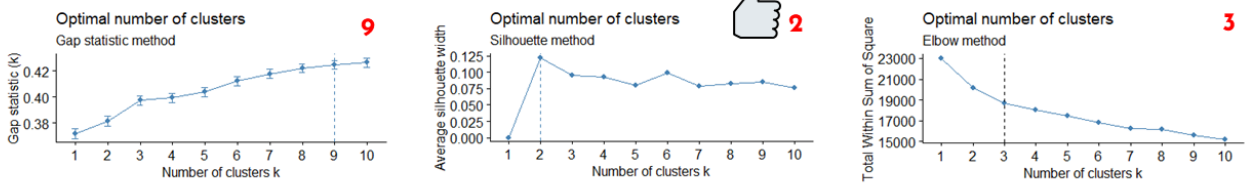
7.1 METODO NO JERARQUICO

Para poder realizar una segmentación de clientes con este método, primeramente, debemos hallar el número de clústeres deseados para aplicarlo al modelo.

Transformación de variables para realizar clustering

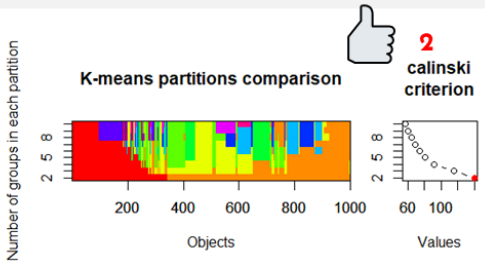
- Conversión de variables categóricas a dummys
- Datos escalados para contrarrestar grandes diferencias entre los valores de las distintas variables

El primer paso a realizar es definir el número adecuado de clústeres que queremos obtener, para ello, podemos aplicar distintas técnicas que nos proporcionan resultados que pueden ayudarnos a definirlo, ya que depende de distintas interpretaciones y análisis.



Para determinar el número de K óptimo, he utilizado 4 métodos gráficos como son "Elbow", "Silohuette", "Calinski", "Gap Statistic". Dado que los resultados son dispares, he recurrido a un Grid a través de la función nbclust() usando 17 índices y 3 métodos distintos, ofreciéndome 51 resultados diferentes, en donde predomina con diferencia K=2 como número de clústeres recomendado.

Número de K	1	2	3	5	6	7	8	9	10
Resultados	1	22	8	3	3	2	4	1	7



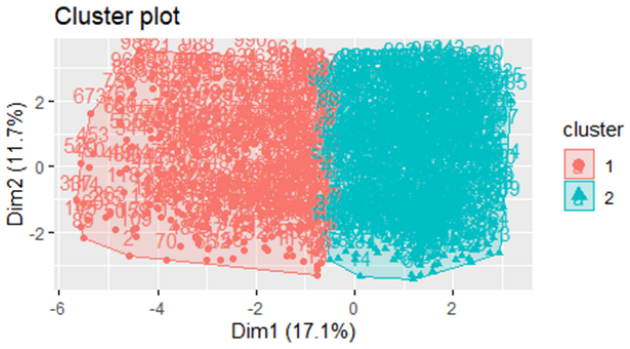
7.1.1 K-MEANS

Para la realización del modelo, se usa la función kmeans() de la librería "stat", el valor k=2 hallado anteriormente, configurándolo para que realice 25 iteraciones. Se aplica con el conjunto de datos escalados ("resultsca").

Variable	Clúster 1 (658)	Clúster 2 (342)
City	Yangon 35%	Naypytaw 34,8%
Customer	Normal 51,1%	Member 52,3%
Gender	Male 51,7%	Female 53,5
Payment	Cash 34,5%	Ewallet 35,4%
Product	Fashion 18,2%	Sports 17,8%
Day	Domingo 15,5% Miércoles 15,2%	Domingo 18,1% Miércoles 17%
Month	Marzo 36,3	Enero 39,5%
Week	4ª 9,4%	4ª 10,8%
Hour	13h 10,8% 10h 10,6%	19h 13,7% 14h 9,4%
Rating	7,01	6,9
Tmed	27	26,5
Total	163,12$	577,9$
Quantity	4 uds	7,88 uds
Price	40,57$	77,3$

En la tabla de la izquierda podemos observar las siguientes características más llamativas entre grupos de clientes:

- La mayor diferencia se halla en el Total de ingresos obtenidos, las cantidades medianas realizadas en cada venta, y el precio mediano pagado, siendo con diferencia superior en el Cluster 2
- Los resultados de las variables están distribuidos muy equitativamente. Parece que el Cluster 2 tiene rasgos un poco más marcados, cliente mujer, compra en domingo o miércoles, principalmente en enero, entorno a las 19:00
- El cluster 2 contiene 34,2% de las observaciones (clúster 1 tiene 65,8%), por lo que parece que está bien definido, y son aproximadamente 1/3 de las ventas las que más ingresos generan

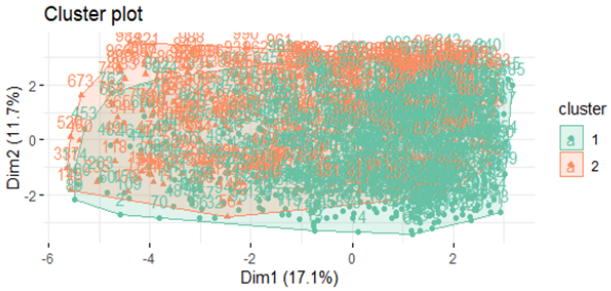


7.1.2 PAM

Para la realización del modelo, se usa la función pam() de la librería "cluster", el valor k=2 hallado anteriormente y el resto de parámetros se dejan por defecto. Se aplica con el conjunto de datos escalados ("resultsca").

Variable	Clúster 1 (670)	Clúster 2 (330)
City	Mandalay 39%	Yangon 54,8%
Customer	Normal 58,1%	Member 66,7%
Gender	Male 58,7%	Female 67,9%
Payment	Cash 39,3%	Ewallet 52,4%
Product	Fashion 19%	Home 19,7%
Day	Domingo 19,1%	Lunes 19,2%
Month	Enero 40,6%	Marzo 46,4%
Week	6ª 11,3%	12ª 15,5%
Hour	19h 10,6% 18h 10,3%	19h 12,7% 13h 11,5%
Rating	7,1	6,6
Tmed	26º	28º
Total	199$	409,89$
Quantity	5 uds	7 uds
Price	48,50$	71,54$

- Con este método podemos observar que las características del perfil de nuestros grupos de clientes están mucho más definidas y diferenciadas. Los datos no se encuentran tan repartidos como con el método anterior.
- Podemos generalizar que el grupo que más nos interesaría en cuanto a ingresos es el 2, siendo mayoritariamente de mujeres de Yangon, pagan mediante Ewallet productos de casa. La compra la realizaron en marzo, los lunes, a las 19h. Generaron unos ingresos medianos de 410$ y con un carrito de 7 productos con precio mediano de 71,54$
- El clúster 1 está formado por 67% de las observaciones, mientras que el clúster 2 se compone del restante 33,3%.
- Estos resultados podrían ser muy útiles para realizar una campaña de marketing, pudiendo enfocarnos muy específicamente en uno de los dos perfiles de clientes obtenidos, según queramos incentivar la compra en el grupo 1 o premiar/fidelizar al grupo 2.

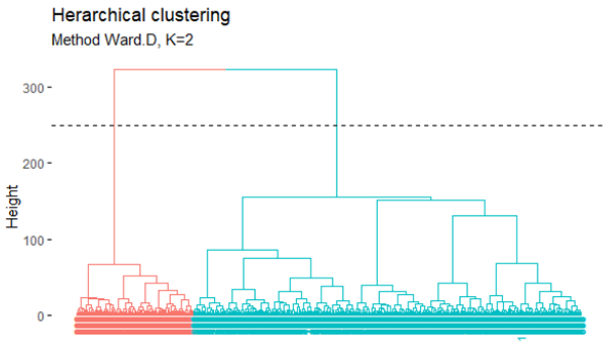


7.2 METODO JERARQUICO
7.2.1 HIERARCHICAL CLUSTERING ANALYSIS

Se usa la función hclust() con el método "Ward.D" de la librería "stats" aplicándose a la matriz de distancias euclídeas de los datos "resultsca". Con el modelo jerárquico primero obtenemos el dendograma y visualizándolo, decidimos que hay una clara división entre 2 grupos distintos que, serán tomados como modelo.

	Cluster 1 (800)	Cluster 2 (200)
City	Yangon 35,4%	Naypyitaw 39%
Customer	Normal 50,7%	Member 56,7%
Gender	Male 50,5%	Female 53,5%
Payment	Ewallet 35,1%	Cash 35%
Product	Fashion 18,5%	Sports 27%
Day	Domingo 16,1% Miércoles 15,4%	Domingo 17,5% Miércoles 17,5%
Month	Marzo 36,6%	Enero 41,5%
Week	6ª 9,4%	3ª 12%
Hour	19h 11,2% 13h 10,4%	10h 12,5% 11h 12%
Rating	7	6,65%
Tmed	27º	26,4º
Total	195$	706,9$
Quantity	1 ud.	10 uds
Price	47,33$	79,90$

- Esta segmentación realizada destaca por ser la que hace una división más clara entre el grupo 1 que genera menos ingresos y vende menos cantidades y el grupo 2 que es justamente lo contrario.
- El resto de datos al igual que en la segmentación K-means, se encuentran distribuidos equitativamente entre cada una de sus categorías. Solamente destaca con una proporción algo mayor la característica de miembro entre quienes más ingresos generan.
- Las observaciones se encuentran divididas entre el 80% del grupo 1 y el restante 20% del grupo 2.



6.3 SELECCION MEJOR CLUSTER

Con los resultados obtenidos de los modelos anteriores, he optado por escoger el modelo PAM. Este nos va a permitir realizar una campaña de marketing hacia un segmento de nuestros clientes más concreto y determinado. Las características que nos ofrece este modelo de cada uno de los dos grupos, nos permiten enfocarnos en nuestros clientes en función de los ingresos que generan, bien para incentivar su compra o para premiarla/fidelizarla.

Los otros 2 modelos K-means e Hierarchical Cluster, mostraban también una división clara entre un grupo que generan más ingresos y otro grupo que genera menos. Pero estos resultados, no nos permiten tener una idea clara de que tipo de perfil de cliente es el que realiza la compra, algo importante para nuestro objetivo.

Una vez seleccionado el modelo PAM como el más útil, procedo a realizar predicciones sobre nuestros datos, para comprobar la capacidad de hallar a qué grupo pertenece cada observación de nuestro conjunto de datos.

"El poder tener la capacidad de determinar a qué grupo pertenece cada cliente, podría ayudar en un caso práctico como el siguiente: el cliente va a realizar una compra y en el momento del pago, obtendría un descuento, en función del segmento al que pertenezca, teniendo en cuenta todas las variables de su proceso de compra"

8. PREDICCIÓN

8.1 MODELO PREDICTIVO

Una vez seleccionado nuestro mejor modelo de clustering (PAM), pasamos a la predicción en nuestro conjunto de datos de cada una de los dos grupos. Recordemos que el grupo 1 estaba formado por un 66.7% de las observaciones y el grupo 2 por el 33% restante. Esto nos permitirá utilizar modelos predictivos binarios.

Para realizar las predicciones, el conjunto de datos se dividirá de la siguiente manera:

A continuación, realizamos algunas tareas de "Featuring Engineering" sobre el dataset:

Para la optimización de los resultados, en algunos de los modelos, las variables correlacionadas son eliminadas.

Para realizar las predicciones usaremos distintos tipos de modelos, que con diferentes parámetros configurables nos permitirán optimizar sus resultados (aplicado en métodos que lo permitan). Para ello utilizaremos "Grids" que nos permitan ver cuáles serían los resultados, según cada configuración. A su vez los resultados obtenidos por cada método podrán ser evaluados y comparados con distintas métricas creadas y agrupadas en una tabla.

La mayoría de métodos utilizados son de clasificación, excepto el modelo GLM en el que se aplica una regresión y los valores mayores de 0.5 son considerados como 1 y los menores como 0.

A continuación, se muestra un listado de los modelos utilizados, el tipo de objetivo, las librerías utilizadas y sus parámetros con los valores utilizados para las predicciones.

Modelo	Método	Tipo	Librerías	Parámetros de ajuste
k-Nearest Neighbors	knn	Classification,		K=2
Random Forest	ranger	Classification,	e1071, ranger, dplyr	N.trees : 100,150,200,500,750,2000 Node.size : 3-9 Sampe_size : 0.55, 0.632, 0.7, 0.8 Splitrule= Gini
Stochastic Gradient Boosting	gbm	Classification,	gbm, plyr	n.trees : 100, 250, 500, 750, 2000 interaction.depth : 1, 3, 5 shrinkage : 0.01, 0.1, 0.3 n.minobsinnode : 5, 10, 15 bag.fraction : 0.65, 0.8, 1
Generalized Linear Model	glm	Regression		
Support Vector Machines	svmRadial	Classification,	kernlab	Radial Basis Function Kernel
Naive Bayes	naive_bayes	Classification	naivebayes	laplace
eXtreme Gradient Boosting	xgbTree	Classification,	xgboost, plyr	Nrounds : 1000 max_depth : 2, 4, 6, 10 eta : 0.05, 0.1, 0.2, 0.5, 1 gamma: 0.1, 0.3 colsample_bytree: 0.3, 0.5, 0.7 min_child_weight : 1, 3, 5, 7 subsample : 0.25, 0.5, 0.75, 1
Random Forest	rf	Classification,	randomForest	Mtry=4 Ntree=500
Bagged AdaBoost	AdaBag	Classification	adabag, plyr	Mfinal=500 Boos = True
Linear Discriminant Analysis	lda	Classification	MASS	
CART	rpart	Classification	rpart	Minsplit = 5, 10, 15

Gracias a la información obtenida de los modelos Cart, Adabag, Random Forest, Ranger, xgbTree y GBM, apreciamos que las variables más importantes son:

- Tmed
- Tax5
- Daynum
- Cogs

Para poder evaluar los resultados obtenidos, nos hemos basado en los siguientes parámetros obtenido de cada modelo:

- Matriz de confusión
- Errores de clasificación
- Accuracy: porcentaje total de acierto
- AUC (Area Under Curve)
- Precision: es la proporción de los datos clasificados de un grupo, que realmente lo son. Un valor alto indica que hay pocos falsos positivos.
- Recall: es la proporción de los datos de un grupo que realmente se clasifican como tal. Un valor alto indica que hay pocos falsos negativos.
- F1 Score: un valor alto cercano a 1, indica una perfecta precisión y recall.
- Otros como "Mean Decrease Gini", "RMSE", "pvalue" y "Kappa"

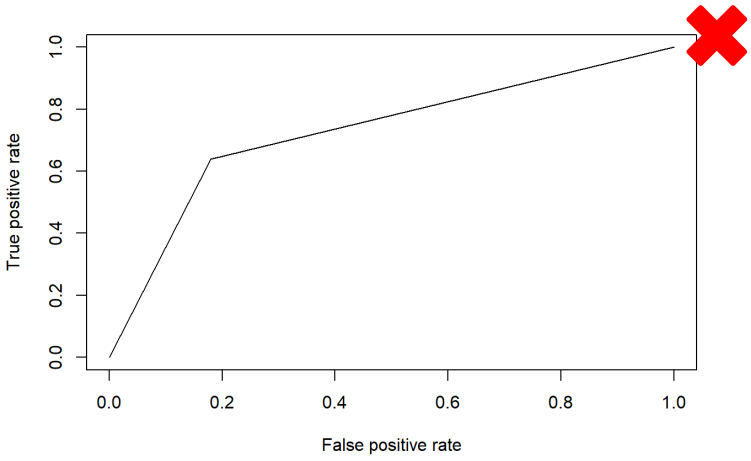
En la tabla siguiente, podemos ver los resultados de nuestras predicciones, ordenados de mayor a menor Accuracy, y donde se recogen el resto de valores utilizados para la evaluación:

Modelo	Accuracy	AUC	Precision1	Precision2	Recall1	Recall2	F1grupo1	F1grupo2	ErrorClas
glm	0,98	0,98	0,98	0,99	1,00	0,94	0,99	0,96	6
lda	0,96	0,96	0,97	0,95	0,98	0,92	0,97	0,93	11
svm	0,94	0,93	0,96	0,90	0,96	0,89	0,96	0,90	17
gbm	0,94	0,91	0,97	0,84	0,94	0,92	0,96	0,88	19
xgb	0,93	0,92	0,94	0,90	0,96	0,85	0,95	0,88	21
randomFor	0,91	0,86	0,96	0,77	0,92	0,88	0,94	0,82	28
adabag	0,90	0,88	0,93	0,83	0,94	0,82	0,93	0,83	29
ranger	0,90	0,85	0,95	0,76	0,91	0,85	0,93	0,80	31
NaiveBaye	0,85	0,83	0,88	0,78	0,91	0,71	0,89	0,74	45
rpart	0,80	0,76	0,85	0,66	0,87	0,63	0,86	0,64	61
knn	0,77	0,73	0,86	0,58	0,82	0,64	0,84	0,61	69

Los mejores resultados son los ofrecidos por el modelo GLM, LDA, SVM Y GBM. Observamos como los dos primeros tienen una alta precisión y un muy buen desempeño a la hora de clasificar cualquiera de los dos grupos. Sin embargo, en el tercer y cuarto modelo, se comienza a apreciar una cierta dificultad a la hora de clasificar el grupo 2.

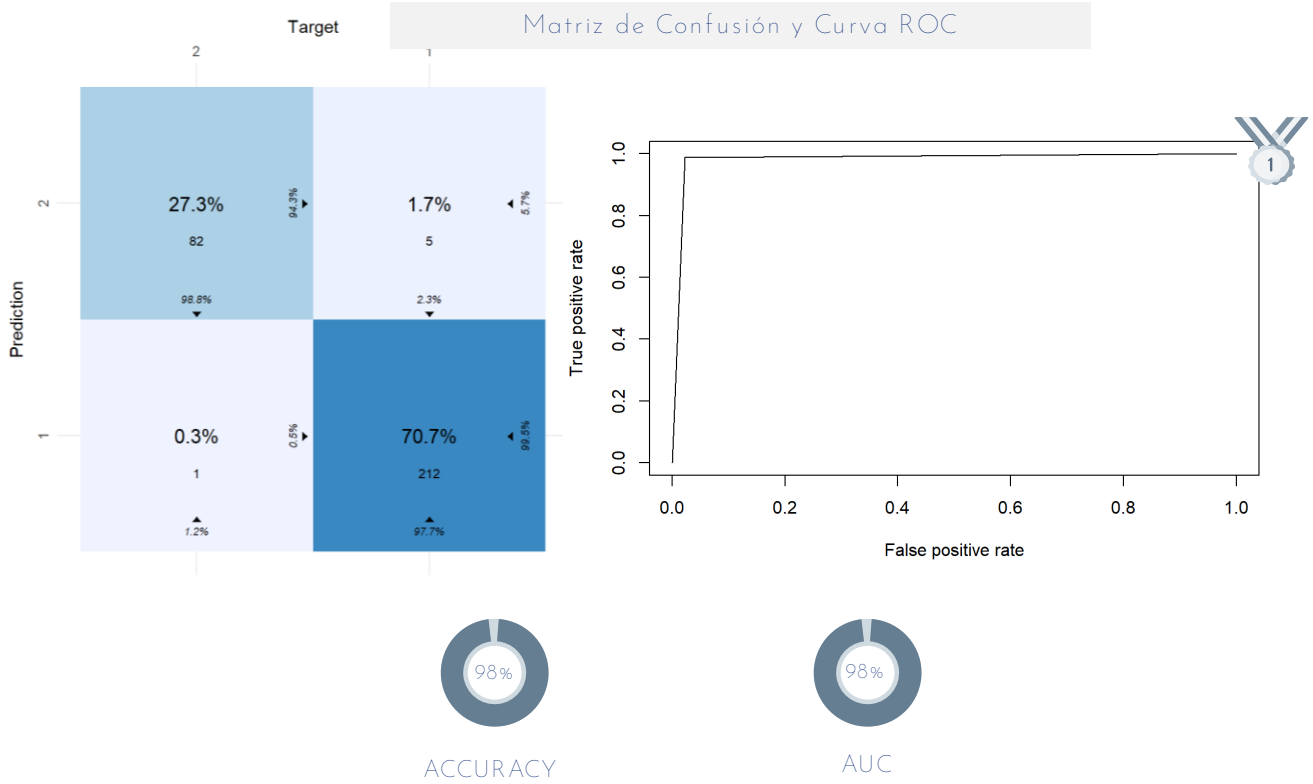
Es un resultado muy bueno, y que destaca muy claramente si lo comparamos gráficamente con el peor de los resultados con el modelo KNN.

Matriz de Confusión y Curva ROC
Modelo KNN



- Tan solo obtiene un 77% de precisión, como se aprecia en la matriz de confusión si sumamos los aciertos al clasificar los datos en el grupo 1 (59.3%) y el grupo 2 (17.7%)
- El Área Bajo la Curva (AUC) indica que este modelo ofrece solo un 73% de probabilidad para poder distinguir entre un grupo y otro.
- El F1 Score indica que el grupo 1 es clasificado notablemente bien (84%), sin embargo, clasificando al grupo 2, encuentra bastante dificultad (61%) y no es muy fiable.

MEJOR MODELO GLM



Sin lugar a duda, nuestro mejor resultado es con el modelo GLM:

- Accuracy del 98% con apenas 6 errores de clasificación de los 300 datos de test.
- Con valores de Precisión del 99%, Recall 94% y F1 Score 96% para el grupo 2, se indica que a pesar de que los datos de estos grupos estaban en inferioridad clara frente a los del grupo 1, los modelos han podido resolver la dificultad para predecirlo con una fiabilidad muy alta.
- Un AUC del 98% nos indica casi un modelo perfecto a la hora de distinguir entre cada grupo.

Una vez realizado el último procedimiento, podemos decir que contamos con un grupo de clientes bien identificado, y que puede ser predicho con una alta fiabilidad. Ahora es cuando hay que dejar paso a otros especialistas, para que utilicen la información proporcionada de la manera correcta. Es posible que muchos departamentos como Marketing o Ventas, se froten las manos con este tipo de información, ya que, si son bien utilizados, pueden ayudar a mejorar los resultados de muchas compañías.

9. CONCLUSIONES

01 A pesar de la sencillez del conjunto de datos, se pueden realizar múltiples análisis y obtener distintas conclusiones según nuestros intereses y objetivos.

02 Aplicar "Feature Engineering" es fundamental a la hora de generar valor en el dataset, algunas de las variables más importantes para los modelos provienen de ello, como Daynum y Tmed.

03 La elección del modelo de predicción o de clustering adecuado es muy importante, ya que condiciona y afecta a nuestros resultados.

04 No hay un solo camino para realizar todo el proceso, continuamente se realizan modificaciones y mejoras y se debe volver a puntos anteriores.

05 Con los análisis realizados, es fácil concluir que no es un conjunto de datos real, si no, preparado para el aprendizaje

06 ¿Qué habría sucedido con datos reales? ¿Cómo sería la evolución a lo largo de un periodo de tiempo mayor? ¿Qué otras variables que no hemos tratado pueden estar influyendo en los clientes?

Ha sido duro y ha supuesto un esfuerzo extra. No es fácil compatibilizar el trabajo diario, con el Máster, y más en un año tan complicado con el COVID-19 en nuestras vidas. Pero ha merecido la pena, y espero poder aplicar en corto plazo todos los conocimientos adquiridos.

MUCHAS GRACIAS A

Los profesores por la dedicación y el esfuerzo mostrado a lo largo del curso por enseñar y hacernos aprender/comprender. También por los comentarios motivadores y didácticos en los resultados de las distintas tareas.

A mi familia y en especial a mi pareja Perla, por la paciencia y el apoyo proporcionado, incluso cuando no me he podido dedicar a ellos como se debe, por falta de tiempo.

Juan Manuel Ortiz Ramírez

10. BIBLIOGRAFÍA Y ANEXOS

*Dataset
https://www.kaggle.com/aungpyaeap/supermarket-sales

*Ideas Big Data en Retail
https://cleverdata.io/big-data-retail/
https://cleverdata.io/basket-analysis-machine-learning/
https://www.cognodata.com/blog/habitos-consumo-machine-learning-evolucion-retail

*Páginas de referencia
https://stackoverflow.com/
https://www.r-graph-gallery.com/index.html
https://github.com/

*"The Caret Package" por Max Kuhn
https://topepo.github.io/caret/index.html

*"R para ciencia de datos" por Hadley Wickham & Garrett Grolemund
https://es.r4ds.hadley.nz/

*Información Meteorológica
https://www.ncdc.noaa.gov/cdo-web/
https://sy.freemeteo.com/

*Rpubs:
"Modelos de Clasificación" https://rpubs.com/rdelgado/397838
"Introducción a la Graficación con ggplot2" https://rpubs.com/rdelgado/429190

*Material de aprendizaje aportado del Master en Big Data y Business Analytics 19-20

Los anexos incluidos son los siguientes:

- Documento README con instrucciones orientativas
- Conjunto de datos original
- Código en formato RMarkDown y HTML, con comentarios e información técnica más ampliada.
- Script con el proceso de la varible Temperatura
- Archivos generados para aligerar la ejecución de procesos.
- Video con la presentación
- Kernel en Kaggle
- https://www.kaggle.com/juanmick/supermarket-clustering-and-prediction-clusters

*Imágenes e iconos obtenidas y creadas por Flaticon, Freepick y Slidesgo