CODER HOUSE

DOCUMENTO EJECUTIVO DEL MODELO

Integrantes:

María Victoria Yaconisi

Mateo Aguiar

Nicolás Lopez

Mauro Rodriguez Vidal

Fecha entrega:

13/04/2022

Nombre y nro. grupo:

Cafeinomanos - Equipo 1

Comisión nro.:

22740

Profesor:

Luca Citta

Tutor:

Jorge Ignacio Lara Ceballos

Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal

ÍNDICE

Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal

VERSIONADO

Fecha	Descripción	Nro. versión
01/02/2022	- Se añade: I. Presentación de la empresa. II. Preguntas y objetivo de la investigación. III. Conformación del equipo de trabajo. IV. Análisis de los componentes principales. V. Indicadores de la fuente del dataset y los criterios de selección (Data Acquisition). VI. Generación del primer Data Wrangling y EDA, apuntando a sus datos (insights) hacia análisis univariados, bivariados y multivariados. VII. Contar la historia de los datos. VIII. Filtros aplicados a los datos y dataset final para analizar. IX. Plantear objetivos para esos datos. X. Detección de outliers.	**Versión 1**
24/02/2022	- Se añade: I. Elección de variable target. II. Preparación de los datos. III. Selección de algoritmos candidatos con la definición de parámetros a probar para cada algoritmo, para el proceso de entrenamiento del modelo.	**Versión 2**
17/03/2022	- Se añade: I. Comparación de los modelos realizados con presentación de distintas medidas y determinar cuál es el mejor con el fin de hallar primeras conclusiones.	**Versión 3**
13/04/2022	- Desarrollo del documento ejecutivo del modelo. - Arreglos grales y cierre archivo .ipynb (Jupyter Notebook).	**Versión final**

Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal

DESCRIPCIÓN DEL CASO

Somos una organización ad honorem, creada por 4 integrantes fanatizados y motivados por el café. El nombre de la misma es 'Cafeinomanos' y nos dedicamos a brindarles soluciones en cuanto a lo que la ciencia de datos abarque y esté al alcance, a distintas empresas, sobre todo aquellas relacionadas con el ambiente y mundo del café, con el fin de resolver ciertas necesidades haciendo nada más y nada menos que hablar a los datos.
El corriente caso abarca resolver un interrogante importante para una cadena de café, con el sentido de poder orientar sus campañas de marketing y asegurarse que las mismas estén bien orientadas hacia un nicho en particular de su clientela.
La cadena de café nos brindará ciertos dataset de los cuales partimos para nuestro análisis y a través de los cuales intentaremos resolver nuestra premisa principal.

El análisis se encuentra detallado a lo largo de 3 archivos .ipynb (Jupyter Notebook) y los mismos además se encuentran publicados en nuestro repositorio de github. El primer archivo corresponde a parte de la primer entrega (ver versionado página 2) y comprende desde una breve presentación del caso, detalles de la fuente de datos de la cual partimos, descripción breve de cada dataset y, por último, su correspondiente Data Wrangling y EDA.

Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal

Luego, en el segundo archivo corresponde a los análisis univariado, bivariado y multivariado de los datos, una breve descripción de los mismos y de los filtros aplicados, y como último, la presentación del dataset final a utilizar para el caso.

Finalmente, el tercer archivo comprende desde la elección de la variable target, la preparación de los datos, elección de el o los algoritmos y su eventual aplicación con análisis de los resultados.

Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal

OBJETIVO

El objetivo de la corriente investigación tiene como objetivo analizar, a través de distintos datasets de una cadena de café, sus ventas correspondientes al período de abril 2019 según distintos factores. El hito principal y motivación por parte del cliente es la búsqueda de una acertada campaña de marketing orientando la misma a aquellos nichos y segmentos de clientes aún sin explotar y así atraer a los mismos hacia la empresa. Logrando como resultado un incremento en la totalidad de sus clientes. A partir de las ventas brindadas por parte del cliente sobre abril 2019 ¿A qué nichos y perfiles de clientes necesitarán orientar sus próximas campañas publicitarias de manera que los resultados se vean reflejados en un incremento en ventas?

Para esto, analizaremos el impacto en una variable que toma valor igual a 0 si la persona gasta menos de 20 dólares, y 1 si gasta más.

Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal

DESCRIPCIÓN DE LOS DATOS

Como mencionamos anteriormente, contamos con un conjunto de distintos datasets orientados al café y a lo que su negocio respecta. Los mismos son: Staff, Customers, Generations, Product, Sales Target, Sales Outlet, Pastry Inventory, Sales Recieps (201904) y Dates.

a. Customers ☞ Este csv contiene información referida en cuanto a los clientes con sus datos principales

b. Generations ☞ Este csv contiene información referida en cuanto a los nombres de generaciones a partir de sus años de nacimiento

c. Product ☞ Este csv contiene información referida en cuanto a los productos con sus características principales

d. Sales Target ☞ Este csv contiene información referida en cuanto a los targets establecidos

e. Sales Outlet ☞ Este csv contiene información referida a los sales outlet y algunas características

f. Pastry Inventory ☞ Este csv contiene información referida a los manejos de stock de algunos productos de manera diaria

g. Sales Recieps (201904) ☞ Este csv contiene información referida en cuanto a las ventas ocurridas en el período de abril 2019

Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal

h. Dates ☞ Este csv contiene información referida a fechas, intentando ampliar la información a partir de una fecha determinada

i. Staff ☞ Este csv contiene información referida a los empleados de la cadena y a sus principales datos

El dataset utilizado para la realización de los distintos análisis fue proporcionado por los socios gerentes de la S.R.L. con el objetivo de encontrar los distintos desaciertos, si es que hubiesen, dentro de la campaña de ventas, para así, poder tener un mejor panorama de aquellos verdaderos clientes y de aquellos potenciales.

El mismo se compone de 9 archivos csv (listados previamente) que se encuentran descargados localmente y accedidos mediante la librería de pandas con la función pandas.read_csv para luego ser trabajados uno por uno como distintos dataframes que sus nombres coinciden con el nombre del csv. Consideramos que todos los dataset serán utilizados para así brindar una mejor solución para la investigación

El objetivo de este análisis es identificar los puntos fuertes y débiles dentro de la información que tenemos de las ventas de la sociedad, tomando como base de los datos de las transacciones realizadas en Abril del 2019 de sexo, edad, generación de los clientes, cantidad comprada y precio de los productos dentro de cada orden de compra, horario en que se realizaron las distintas transacciones y los puntos de ventas en donde se realizaron, para

Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal

así, de esa forma, poder generar una campaña de marketing enfocada en potenciar las ventas en un 50% por cada año hasta lograr conseguir el crecimiento deseado por nuestros clientes.

A partir de las limpiezas de los datos, los filtros aplicados y análisis correspondientes, determinamos que los dataframes de: Staff, Dates, Pastry Inventory, Sales Target y Sales Outlet no serán de utilidad para nuestra pregunta. Por lo que decidimos descartarlos.

Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal

HALLAZGOS ENCONTRADOS POR EL EDA Y CONCLUSIONES PRINCIPALES DE ANÁLISIS UNIVARIADO, BIVARIADO Y MULTIVARIADO.

Luego de realizar y aplicar funciones y librerías con intención de efectuar el EDA, el mismo arrojó que es un dataset muy completo. Pudiendo observar así que prácticamente no hay valores NaN ni faltantes.

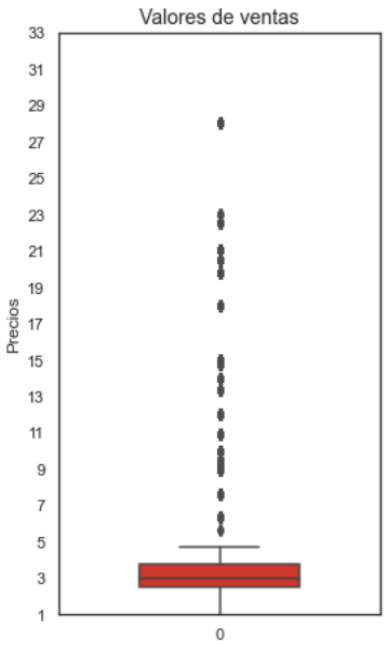
Mencionamos brevemente modificaciones en cada dataset a modo resumen:

a. Customers: se añade un campo de edad, cambiamos las siglas de género por texto. No hay valores NaN.
b. Generations: no hay valores atípicos, ni NaN.
c. Product: eliminamos algunas columnas que eran evidentes que no iban a ser utilizadas a futuro.
d. Sales target: no hay valores atípicos, ni NaN.
e. Sales outlet: la columna de manager contiene un dato NaN. Específicamente el sales_outlet_id = 2.
f. Pastry Inventory: la columna de ‘% waste’ contiene valores ‘0%’ y ‘0’ que corregimos y unificamos en uno.
g. Sales receipts: eliminamos columnas que eran evidentes que no iban a ser utilizadas a futuro.
h. Dates: no hay valores atípicos, ni NaN.
i. Staff: las columnas 7 y 8 no contienen ningún tipo de información. Están vacías, por lo que se precedió a borrarlas. En la columna

'location' algunos campos tienen iniciales y otros números, procederemos a eliminar esta columna directamente.

Análisis univariado: a modo de resúmen, mostraremos algunos gráficos junto con sus conclusiones. A nuestra opinión, los más destacados en cuanto a obtención de conclusiones.

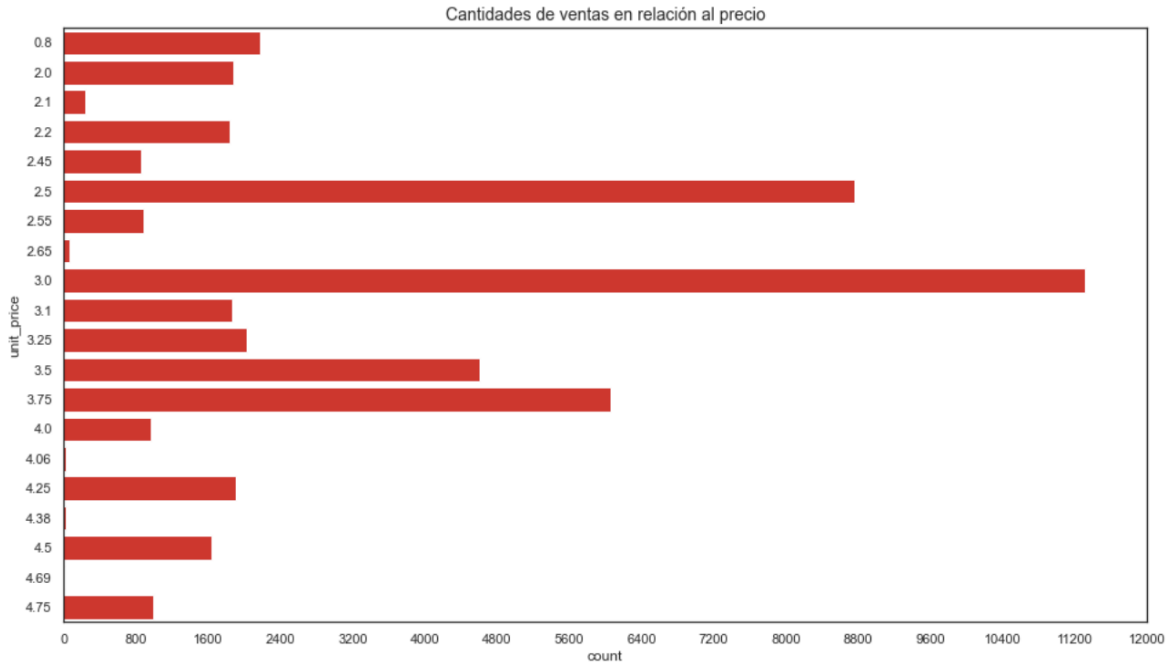
1. Mediante el gráfico de tipo boxplot, también conocido como gráfico de caja y bigotes, se puede ver reflejado los precios de ventas más representativos para Abril del 2019, donde el valor más significativo entre ellos es el de: 3.0



2. En el siguiente gráfico, se reflejan los precios de ventas más representativos de Abril de 2019, para tener una mejor visión de las cantidades más vendidas de productos con relación a su precio. Como se puede apreciar en el siguiente gráfico, los productos con un valor de

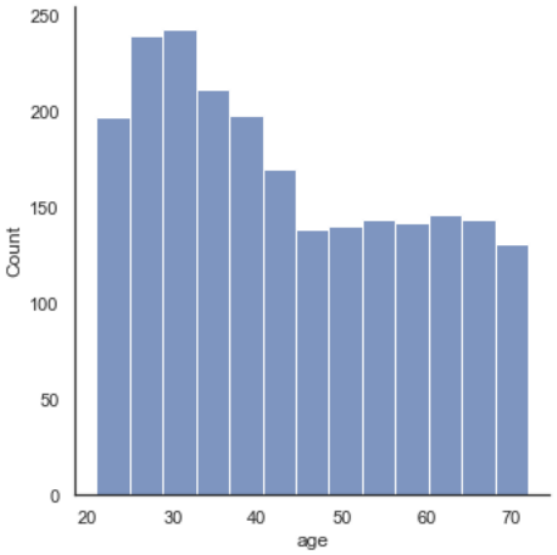
Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal

$2.5 y $3 son los más vendidos en el negocio, los cuales rondan por encima de las 8.500 ventas.



3. Lo que podemos finalmente deducir acerca del análisis hecho sobre la edad de nuestros clientes es que se mantienen un promedio de edad bastante calmado, sin ver picos extremistas en ningún sector de edades en particular. Podría decirse que aumentan un poco aquellos clientes entre 21 y 40 años, pero vemos que la diferencia con las edades que le siguen no son cambios muy abruptos.

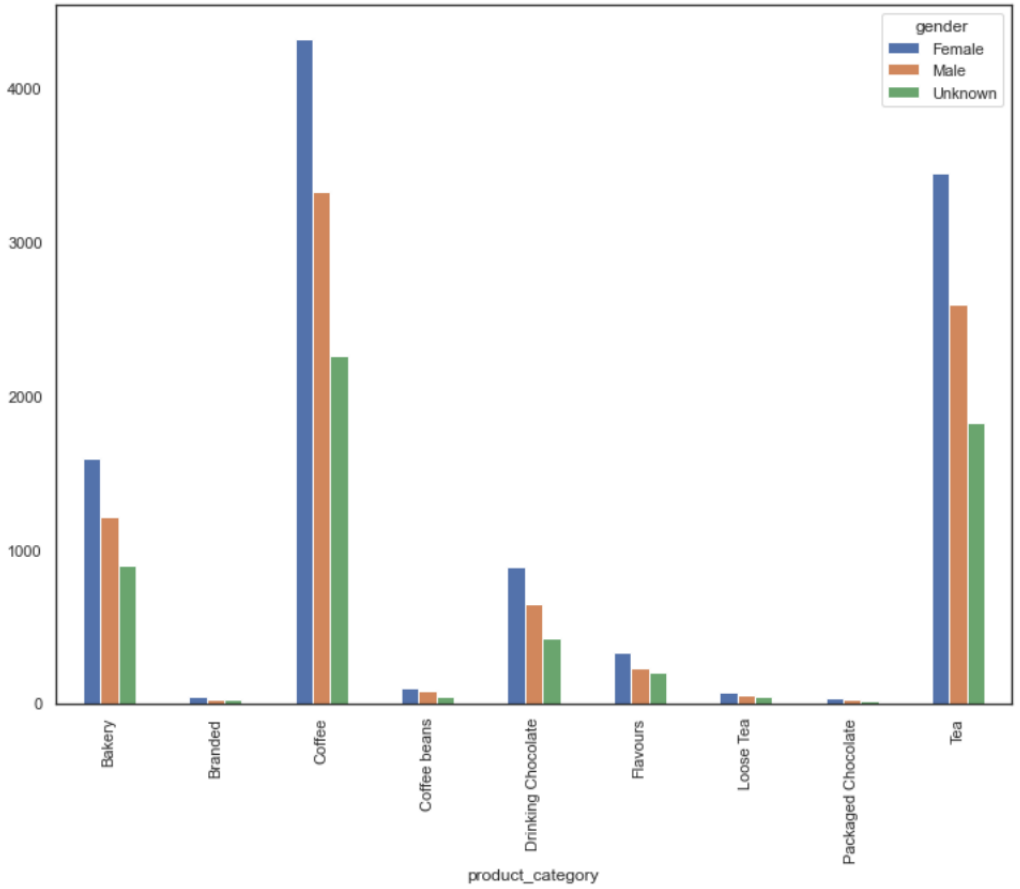
Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal



<u>Análisis</u> <u>bivariado</u>: a modo de resúmen, mostraremos algunos gráficos junto con sus conclusiones. A nuestra opinión, los más destacados en cuanto a obtención de conclusiones.

1. Podemos visualizar que aunque siempre se mantiene que el género femenino tiene mayor cantidad de compras que otros géneros, se mantiene un equilibrio en cada categoría de los productos del café para las compras de abril de 2019

Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal

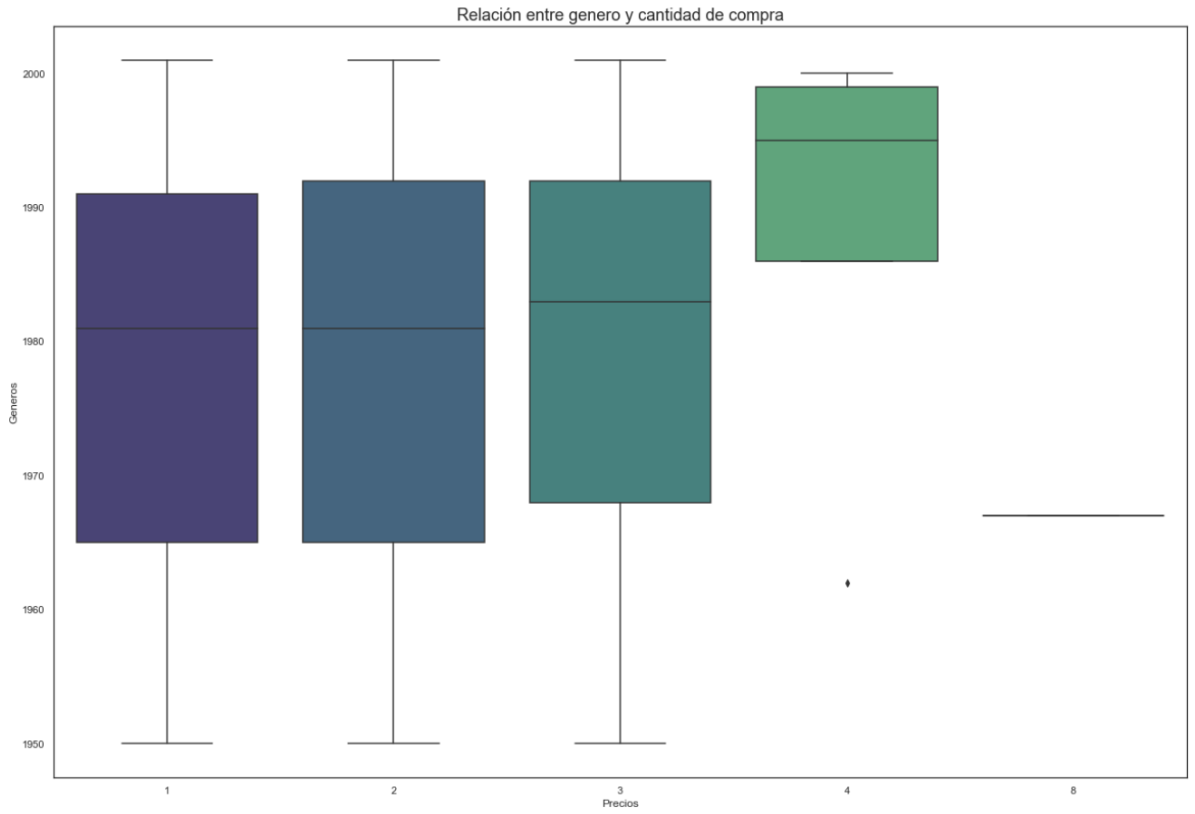


2. En el siguiente gráfico de violín, se puede apreciar que las distribuciones entre hombres y mujeres son muy parecidos, casi idénticos, entre los clientes que nacieron entre 1960 y 1990, también se puede observar que en cuanto al género no definido, observamos que se maneja una media alrededor del año 1995.

Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal

3. Mediante el gráfico de tipo boxplot podemos apreciar las personas nacidas entre 1965 y 1991 son las que realizan compras de 1 a 3 productos por órdenes y son las que mayormente compran, mientras que las personas que mas productos compras, es decir 4 productos por orden manejan una media de nacimiento de 1995, a excepción de los que realizan compras de 8 productos, que son casos excepcionales y son personas nacidas en 1968

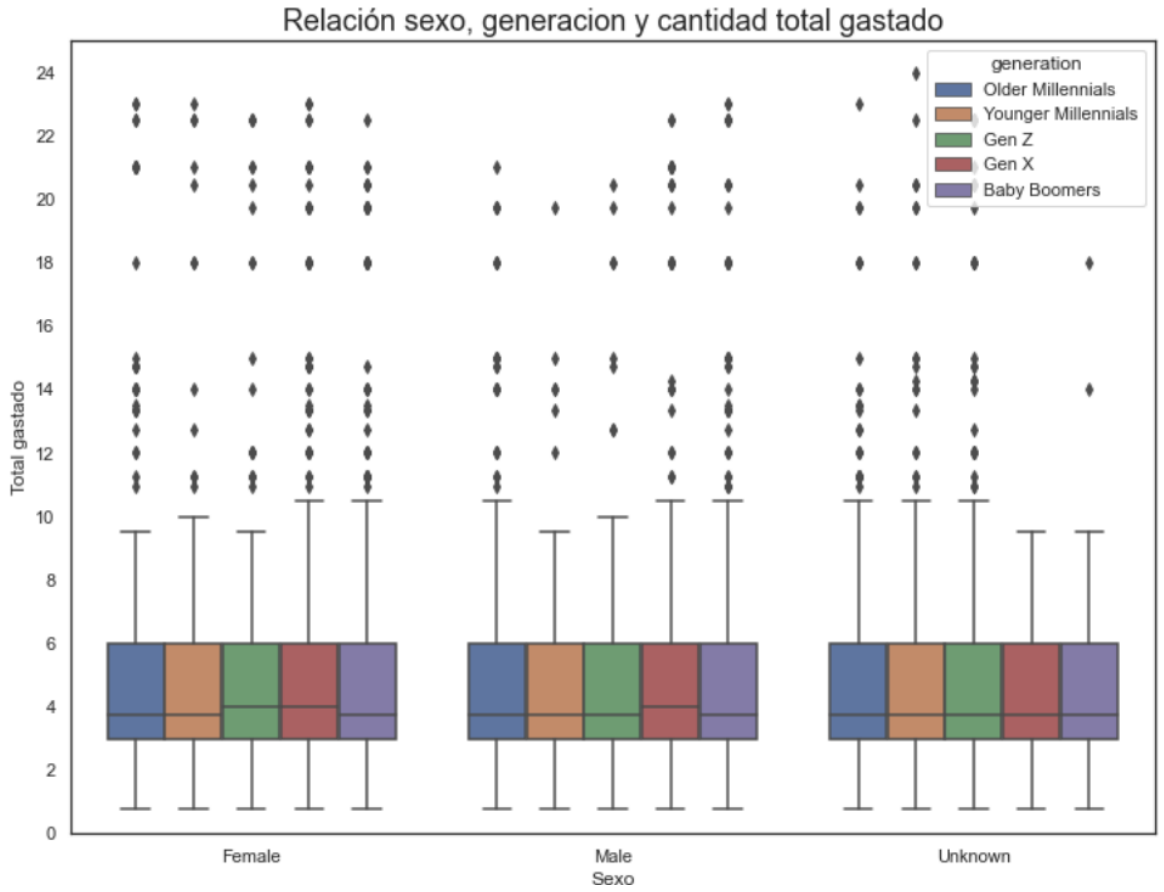
Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal



Análisis multivariado: a modo de resúmen, mostraremos algunos gráficos junto con sus conclusiones. A nuestra opinión, los más destacados en cuanto a obtención de conclusiones.

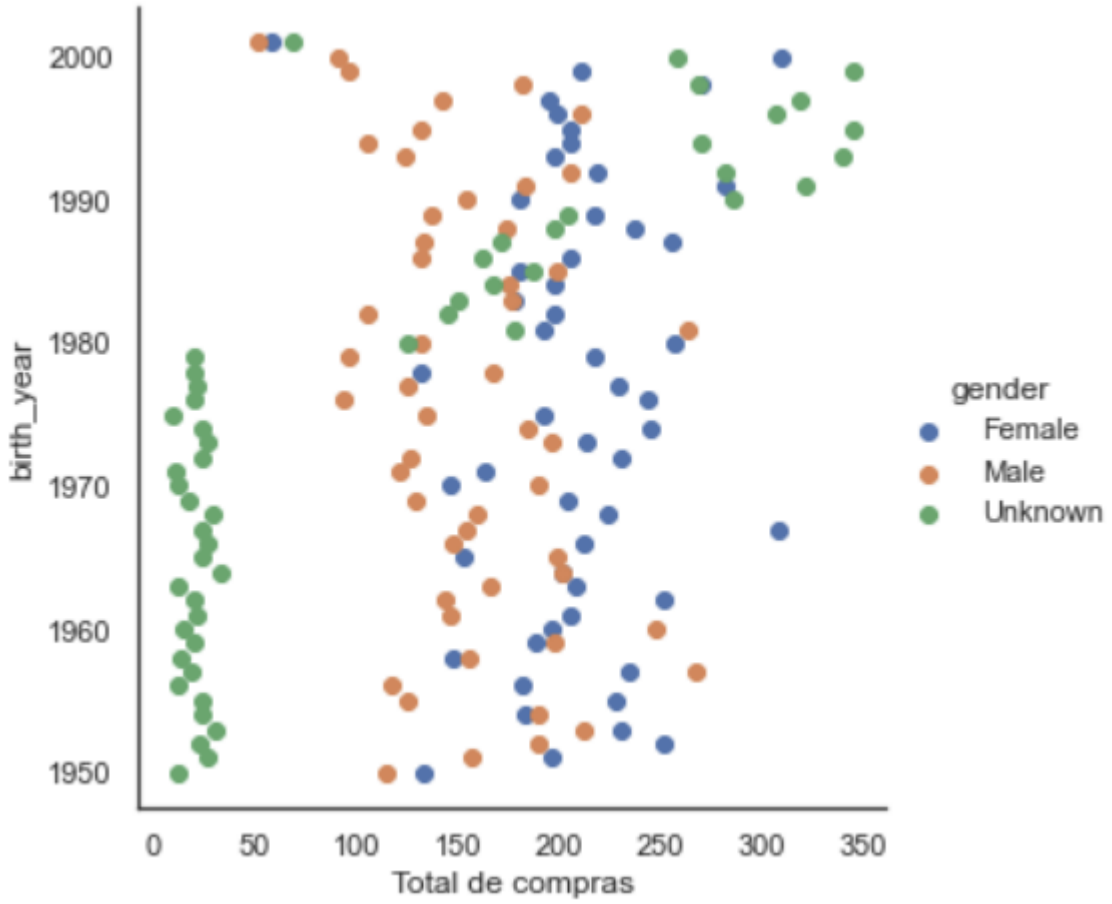
1. Con este gráfico de caja y bigotes podemos apreciar que las distintas generaciones en sus distintos sexos consumen un monto similar, hablando de una media de $3,5 o $4.

Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal



2. Podemos apreciar que la edad y el género de los consumidores no son muy relevantes para poder analizar las ventas, a partir de aquellos nacidos en 1980 de géneros "N" se puede apreciar una mínima correlación entre cantidad de compras y gente más joven.

Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal



Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal

DATASET FINAL

Luego del análisis llevado a cabo en el punto anterior, pasamos a detallar los campos y estructura de nuestro dataset final, el cual utilizaremos como punto de partida para lo siguiente, que es la elección del algoritmo y su correspondiente entrenamiento del modelo, sobre la variable target elegida.

Esta variable target vamos a llamarla 'MoreThan20' y registrará un 0 en aquellos casos donde el total de lo gastado no superó los 20 dólares y 1 cuando sí haya superado o igualado.

Dataset:

```
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   product_id      24685 non-null  int64
 1   quantity        24685 non-null  int64
 2   unit_price      24685 non-null  float64
 3   gender          24685 non-null  int32
 4   age             24685 non-null  int32
 5   generation      24685 non-null  int32
 6   product_group   24685 non-null  int32
 7   tax_exempt_yn   24685 non-null  int32
 8   promo_yn        24685 non-null  int32
 9   new_product_yn  24685 non-null  int32
 10  MoreThan20      24685 non-null  int32
```

	product_id	quantity	unit_price	gender	age	generation	product_group	tax_exempt_yn	promo_yn	new_product_yn	MoreThan20
0	52	1	2.5	2	39	0	5	1	0	0	0
1	52	1	2.5	1	39	0	5	1	0	0	0
2	52	1	2.5	1	39	0	5	1	0	0	0
3	52	1	2.5	0	39	0	5	1	0	0	0
4	52	2	2.5	1	39	0	5	1	0	0	0

Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal

ALGORITMO ELEGIDO

Para realizar el análisis del gasto esperado del cliente, se prueban 2 técnicas de aprendizaje automático con el fin de buscar el que mejor desempeño tenga. Estos son Random Forest (RF) y K-Nearest-Neighbor (KNN). Hay que tener en cuenta que la variable Target elegida es si la persona gasta más o menos de 20 dólares en la cafetería.

El RF plantea generar varios árboles de decisión, dando como resultado el árbol más popular, uno que represente a la mayoría de ellos. Con esto, se evita el problema de un árbol de decisión común, que puede tener problemas de overfitting. Los distintos árboles se entrenan con porciones distintas y aleatorias de los datos.

Para este modelo, se decide que el 25% de los datos será apartado para realizar el posterior testeo. Además, únicamente a fines de que los resultados no cambien cada vez que se ejecuta el código, se selecciona una semilla para el RF.

Por otro lado se usa el algoritmo KNN. Este, es de tipo supervisado que sirve también para clasificar o predecir nuevas muestras. La idea central es buscar “vecinos” por cercanía, en una o más dimensiones. El K significa la cantidad de “vecinos” que queremos considerar, y de ahí elige la clase por “mayoría”. Este número debe ser elegido al momento de crear el modelo, y en este caso seleccionamos K=3.

Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal

MÉTRICAS DE DESEMPEÑO

Para medir y expresar el desempeño de cada uno de los algoritmos, se usa el indicador “F1-Score”. Para llegar a él, se debe mencionar la Sensibilidad y la Precisión.

La precisión se calcula como

Verdaderos Positivos / (Verdaderos Positivos+Falsos Positivos)

Y la sensibilidad se calcula como

Verdaderos Positivos / (Verdaderos Positivos+Falsos Negativos)

Entonces el F1 Score se calcula como:

2 * (Sensibilidad * Precisión) / (Sensibilidad + Precisión).

El algoritmo de Random Forest, en este caso obtuvo un score de 0.9999459839032031 para Training y 0.9988658457550227 para Testing.

Esto da el indicio de que no tenemos Overfitting ni Underfitting.

De todas formas, más adelante se intentarán mejorar estos valores.

Por otra parte, el algoritmo KNN obtuvo un puntaje de desempeño de 0.9953819978935429.

Ambos modelos dan puntajes muy cercanos a 1, por lo que podríamos decir que ambos son muy eficaces a la hora de predecir los tipos de clientes.

Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal

ITERACIONES DE OPTIMIZACIÓN

Buscamos optimizar los hiper parámetros de nuestro modelo ya que estos son los responsables de decidir cómo un modelo puede ajustarse a los datos para producir resultados precisos, para ello utilizamos el método RandomizedSearchCV que nos sirve para buscar la mejor combinación de hiperparámetros posible de un determinado algoritmo. Para ello definimos:

- **n_estimators:** dicho parámetro controla la cantidad de árboles dentro del clasificador.
- **max_features:** Ayuda a encontrar la cantidad de características a tener en cuenta para hacer la mejor división.
- **max_ depth:** Define la altura máxima hasta la que pueden crecer los árboles dentro del bosque. Es uno de los hiperparámetros más importantes a la hora de aumentar la precisión del modelo, a medida que aumentamos la profundidad del árbol, la precisión del modelo aumenta hasta cierto límite, pero luego comenzará a disminuir gradualmente debido al sobreajuste en el modelo.
- **min_samples_split:** Especifica la cantidad mínima de muestras que debe tener un nodo interno para dividirse en más nodos.
- **min_samples_leaf:** especifica la cantidad mínima de muestras que debe contener un nodo después de dividirse.

Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal

- **Bootstrap:** Es un remuestreo estadístico que implica el muestreo aleatorio de un conjunto de datos con reemplazo, permite generar nuevas muestras a partir de una población sin tener que ir a recopilar datos adicionales.
- **Criterion:** Sirve básicamente para cuantificar la calidad del Split y se dividen en 3 criterios importantes:

1. Entropy
2. Gini Impurity
3. Information gain

Donde:

1. Entropy: Representa el orden de la aleatoriedad y ayuda a modelar en la selección de características para dividir, midiendo la pureza de la división en el nodo. Si es igual a 0, significa que es división pura y si es igual a 1, significa una división completamente impura.
2. Gini Impurity: También calcula la pureza de la división en los nodos del árbol de decisión. A diferencia de la entropía, el valor de la impureza de Gini varía entre 0 y 0,5.
3. Information gain: Representa cuánta entropía se eliminó durante la división en un nodo.

Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal

MÉTRICAS FINALES DEL MODELO OPTIMIZADO

Una vez utilizado el método RandomizedSearchCV, podemos evidenciar que nuestras métricas finales son un poco superiores en Testing y un poco inferiores en Training a aquellas previas a la optimización de los hiper parámetros, dando como resultado un score de:

- Training: 0.9995678712256252
- Testing: 0.9991898898250162

Y una diferencia negativa de 0,000378112677578035 para Training y una diferencia positiva de 0,000324044069993978 para Testing.

Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal

FUTURAS LÍNEAS

Como idea a futuro, de este análisis pueden desprenderse varios proyectos posteriores. Sin embargo, uno de los más interesantes (y posiblemente con más uso potencial) es poder lograr un ida y vuelta de información y sugerencias. La idea fundamental es lograr algún método para que, en base al tipo y atributos del cliente (conocido o desconocido), al ingresarlos al sistema sean devueltas sugerencias para lo que se debe (y lo que NO se debe) ofrecer a la persona para maximizar el consumo y gasto.

Entendemos que sería muy interesante poder obtener de manera casi instantánea sugerencias de productos que ese tipo de persona está propenso a comprar, y sugerencias de productos que no debemos ofrecerle para no ahuyentarlo. Es posible que de existir esta suerte de “programa” (que claramente tendrá de fondo modelos de machine learning, y que aprendería con la experiencia), los ingresos de la cafetería aumenten de manera sustancial. Se podría lograr, al conocer las probables preferencias del individuo, una experiencia personalizada que aumente tanto el bienestar del cliente (estaríamos ofreciéndole lo que le gustaría consumir) como las ventas.

Esta idea está inspirada en lo que muchas plataformas y servicios que hoy día consumimos hacen (Redes Sociales, páginas web), pero que también se hace de forma “humana” en muchos negocios. Los buenos vendedores se diferencian de los malos vendedores en gran parte por saber interpretar lo que el cliente necesita. Y esto usualmente se traduce en satisfacción del

mismo, y en aumento de la facturación. Entonces, si se lograra computarizar esta actividad en los negocios, las posibilidades de éxito existen y no creemos que sean bajas.

Nicolás Lopez
Mateo Aguiar
María Victoria Yaconisi
Mauro Rodriguez Vidal

CONCLUSIONES

Como conclusiones, queremos responder a la pregunta inicial, y de donde salió la variable target. Hemos llegado a algoritmos capaces de predecir con muchísima precisión los gastos de los clientes. En particular, ambos algoritmos son capaces de predecir con más de un 99% de precisión si el cliente gastará más o menos de 20 dólares. Para esta precisión es importante recordar que es una base de datos elaborada artificialmente y es probable que por eso sean tan altos los *scores*. Sin embargo, entendemos que el objetivo fue completado de manera satisfactoria.