ONLINE SHOPPERS INTENTION

Trabajo Práctico Data Science

Descripción breve

Análisis y predicción de la intención de compra de usuarios de una página web

Roca, Martín Gonzalo

Descripción del caso de negocio

El mundo actual presenta un escenario de competencia globalizada en el que resulta fundamental aprovechar las posibilidades que brinda la informática a fin de lograr buenas experiencias de compra a los usuarios, a la vez que se intenta maximizar las ganancias y llegar a la mayor cantidad de clients posibles.

Esto lleva a las organizaciones al estudio permanente de la satisfacción de los clientes y a la búsqueda de nuevos nichos de mercado, entendiendo que el auge que las ventas digitales tuvieron en los últimos años ha sido muy significativo.

Objetivo del trabajo

El presente trabajo busca analizar las sesiones realizadas por los usuarios en una página web de compras on-line, en el período de un año. Buscaremos identificar las variables más importantes que nos ayuden a identificar a los usuarios más interesados en comprar los productos, y a poder predecir dicha intención de compra para así poder realizar distintas campañas que ayuden a recomendar productos similares, o accesorios que puedan resultar de interés, brindando a los usuarios una experiencia más integral y completa, a la vez que genere mayores ventas.

Hipótesis a desarrollar

1. A mayor tiempo de duración en páginas similares mayor probabilidad de venta
2. Mayor ExitRate menor probabilidad de compra
3. Mayor PageValue mayor probabilidad de compra
4. Las ventas crecen los fines de semana y los períodos anteriores a días especiales
5. Hay estacionalidad en las compras
6. La mayoría de las compras las realizan al volver a visitar la pagina

Observaciones iniciales del Dataset

El dataset presenta registros respecto a las sesiones que realizaron los usuarios en una página de compra online de productos y cuenta con 12330 registros y 18 columnas que representan las variables identificadas sobre las acciones realizadas en dichas sesiones. Cada registro corresponde a una sesión distinta.

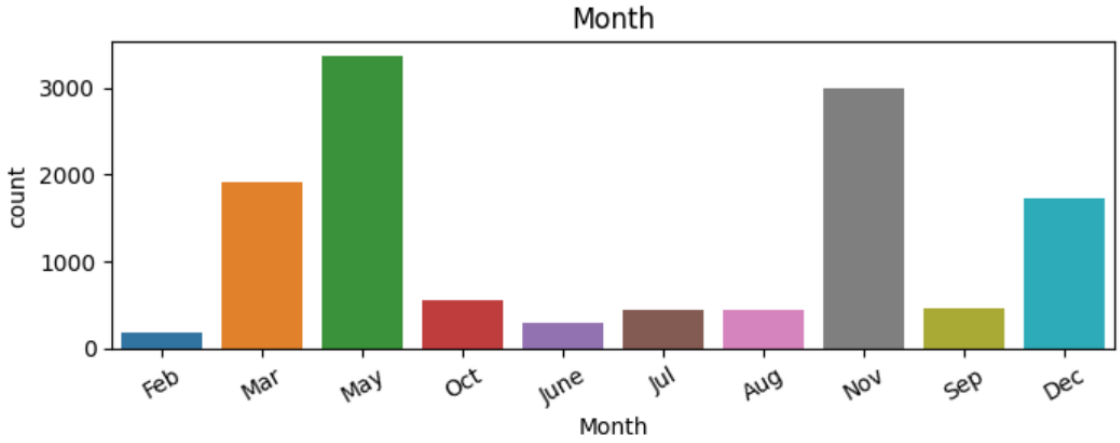
Como primer paso se encontraron 314 registros duplicados y se procedió a su eliminación para comenzar con el análisis de cada variable.

Descripción de las variables:

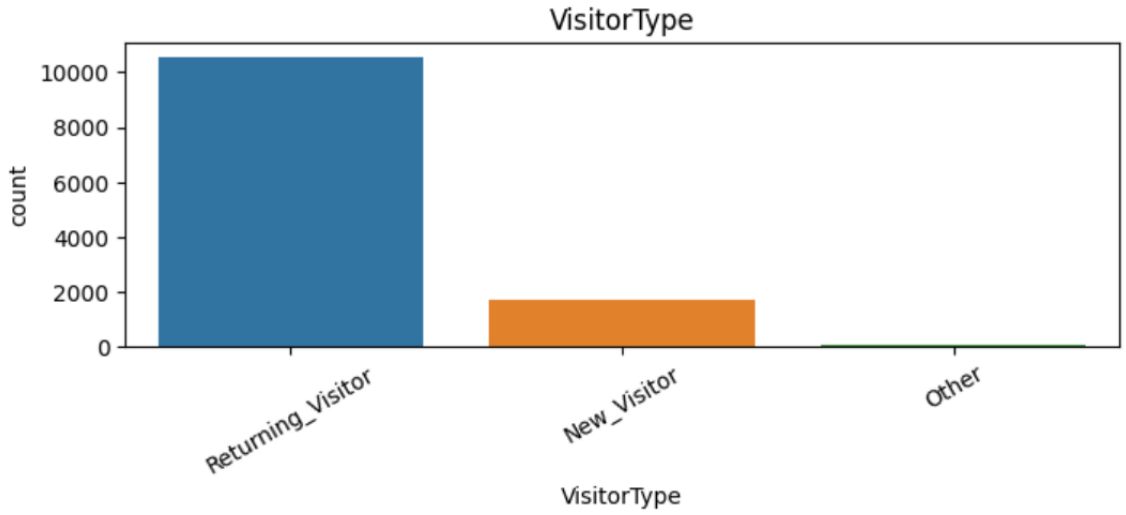
- Administrative: Número de páginas de este tipo visitadas por el usuario.
- Administrative_Duration: El tiempo dedicado en esta categoría de páginas.
- Informational: Número de páginas de este tipo visitadas por el usuario.
- Informational_Duration: El tiempo dedicado en esta categoría de páginas.
- ProductRelated: Número de páginas de este tipo visitadas por el usuario.
- ProductRelated_Duration: El tiempo dedicado en esta categoría de páginas.
- BounceRates: Porcentaje de visitantes que ingresaron a la web a través de esa página y salieron sin realizar ninguna otra actividad
- ExitRates: Porcentaje de vistas a la página que finalizaron en esa misma página.
- PageValues: Valor promedio de la página realizado sobre el valor de la página y/o la finalización de una acción de Ecommerce.
- SpecialDay: proximidad de la fecha de búsqueda a un día feriado o especial.
- Month: Mes de visita a la web.
- OperatingSystems: Sistema operativo utilizado.
- Browser: Navegador web utilizado.
- Region: Region del usuario.
- TrafficType: Tipo de tráfico en la que se categoriza al usuario
- VisitorType: identificador de si el usuario es New Visitor, Returning Visitor, or Other
- Weekend: identificador de fin de semana
- Revenue: identificador de si realizó la compra o no

Análisis Univariado de variables

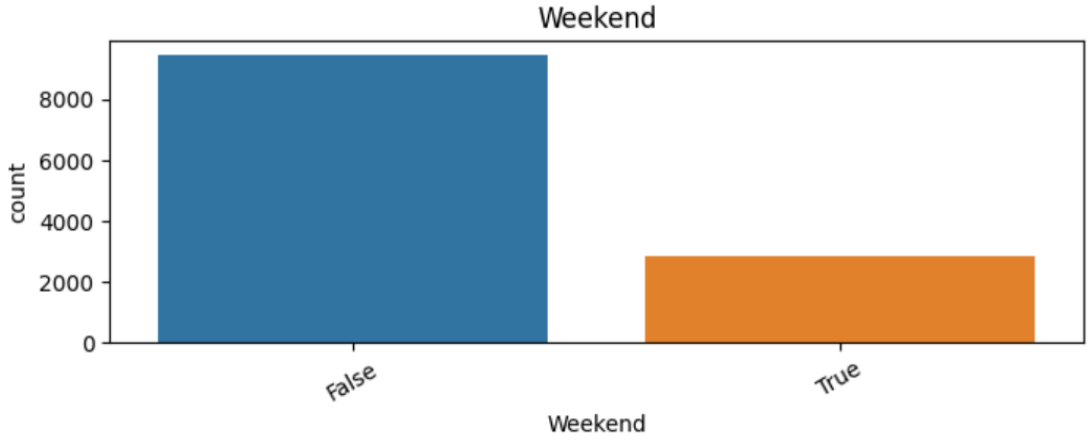
- **Month**: La variable muestra niveles importantes de actividad en los meses de Marzo, Mayo, Noviembre y Diciembre respecto del resto. Además se destaca que no hay datos ni en Enero ni en Abril.



- **VisitorType**: La mayoría de los usuarios son usuarios previos, hay muy poco ingreso de nuevos usuarios, por lo cual sería recomendable fomentar el ingreso de nuevos consumidores a través de alguna campaña.

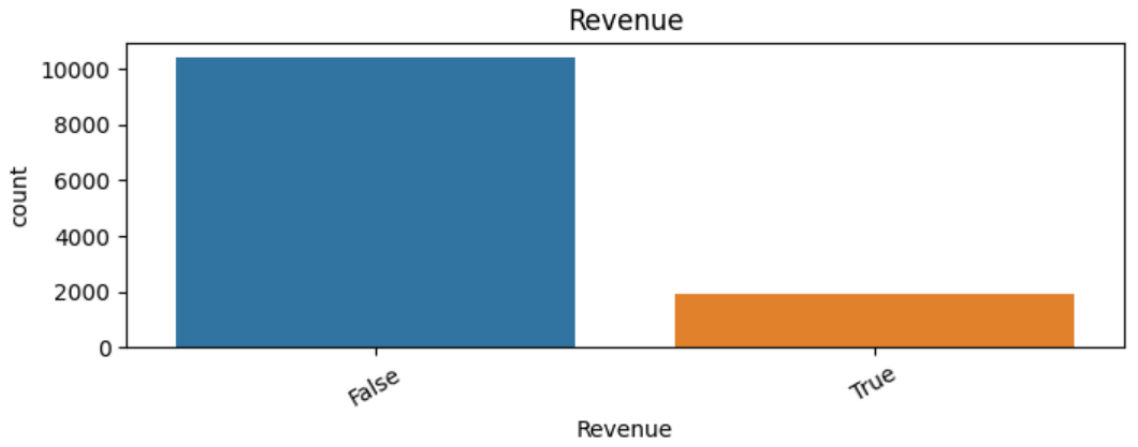


- **Weekend**: No se observa mayor flujo de ingresos a la web durante los fines de semana.



- **Revenue**: Aproximadamente un 20% de las visitas terminan con una compra de producto. Debe considerarse el desbalanceo de clases sobre esta variable al desarrollar el modelo.

Esta será la variable a predecir en los modelos.



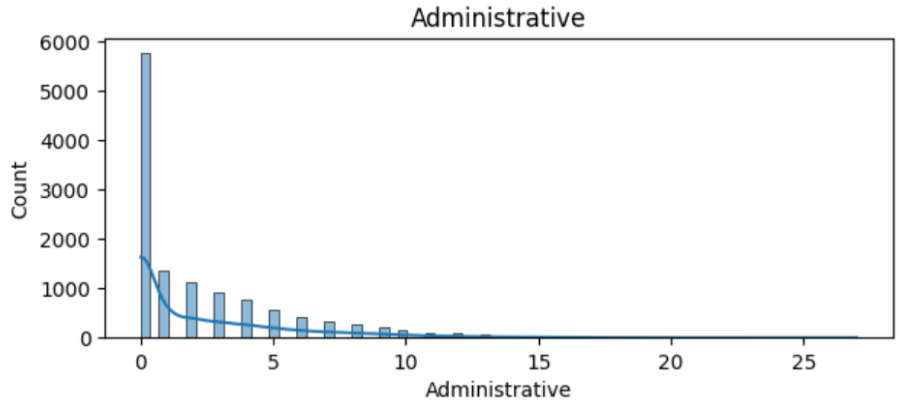
Observaciones estadísticas iniciales sobre variables

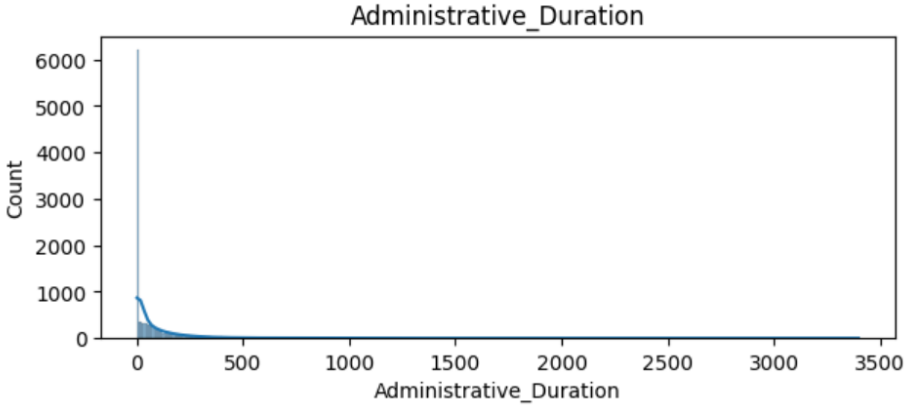
	count	mean	std	min	25%	50%	75%	max
Administrative	12016.0	2.375666	3.343483	0.0	0.000000	1.000000	4.000000	27.000000
Administrative_Duration	12016.0	82.930548	178.584445	0.0	0.000000	11.000000	96.500000	3398.750000
Informational	12016.0	0.516728	1.284001	0.0	0.000000	0.000000	0.000000	24.000000
Informational_Duration	12016.0	35.373225	142.464803	0.0	0.000000	0.000000	0.000000	2549.375000
ProductRelated	12016.0	32.531708	44.772915	0.0	8.000000	19.000000	39.000000	705.000000
ProductRelated_Duration	12016.0	1225.963124	1928.617897	0.0	207.563103	630.413333	1501.208333	63973.522230
BounceRates	12016.0	0.017595	0.039660	0.0	0.000000	0.002632	0.015385	0.200000
ExitRates	12016.0	0.039000	0.042056	0.0	0.013947	0.025000	0.046154	0.200000
PageValues	12016.0	6.043155	18.784765	0.0	0.000000	0.000000	0.000000	361.763742
SpecialDay	12016.0	0.062883	0.201055	0.0	0.000000	0.000000	0.000000	1.000000
OperatingSystems	12016.0	2.124834	0.905283	1.0	2.000000	2.000000	3.000000	8.000000
Browser	12016.0	2.362101	1.711227	1.0	2.000000	2.000000	2.000000	13.000000
Region	12016.0	3.157956	2.403808	1.0	1.000000	3.000000	4.000000	9.000000
TrafficType	12016.0	4.076897	4.008594	1.0	2.000000	2.000000	4.000000	20.000000

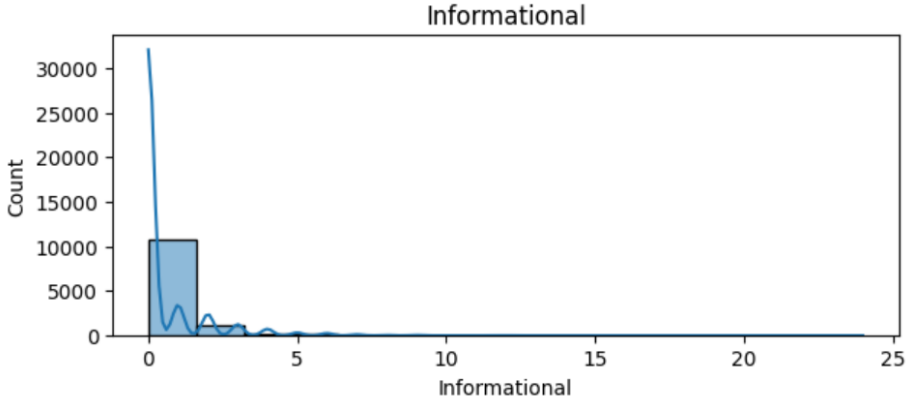
Al observar las descripciones estadísticas iniciales sobre las variables con valores numéricos se observa lo siguiente:

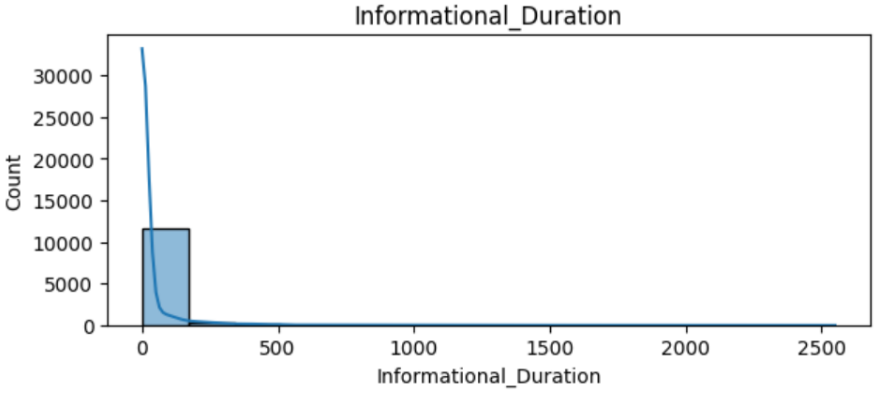
- Diferencia relativa importante entre el valor máximo y el promedio en las variables Administrative, Administrative_Duration, Informational, Informational_Duration, ProductRelated, ProductRelated_Duration. En particular las variables Informational parecen presentar en su mayoría valores cero, ya que se encuentran hasta su tercer cuartil. Deberemos analizar posible tratamiento de Outliers sobre las mismas.

- PageValues y SpecialDay también presentan valores cero hasta su tercer cuartil.

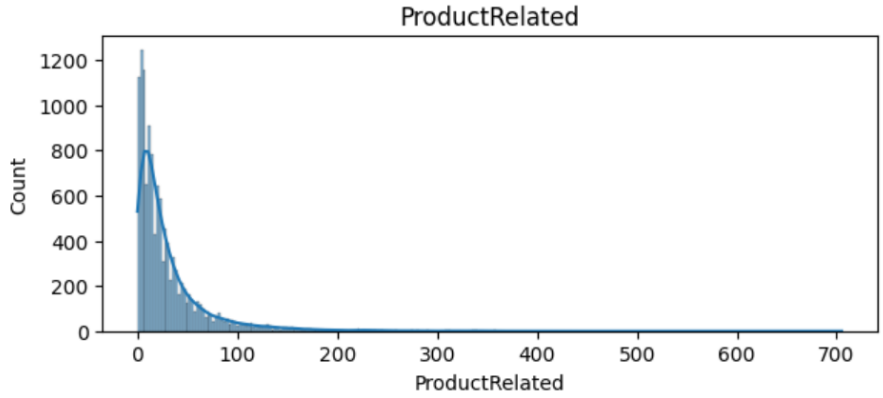

Administrative
Count
Administrative
0
1000
2000
3000
4000
5000
6000
0
5
10
15
20
25

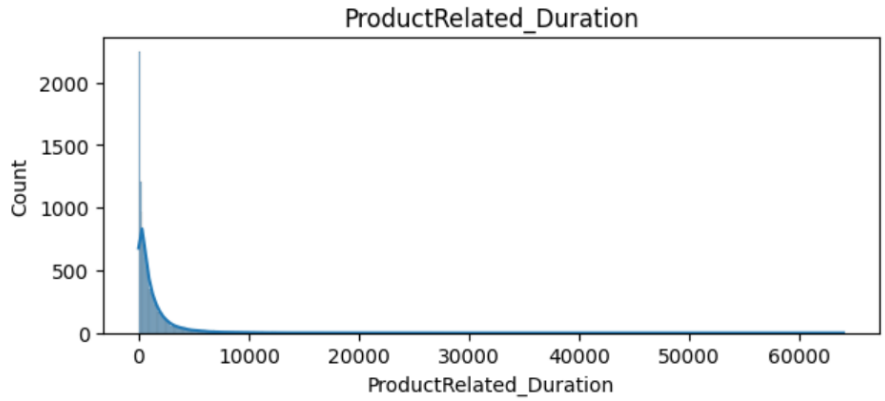

Administrative_Duration
Count
Administrative_Duration
0
1000
2000
3000
4000
5000
6000
0
500
1000
1500
2000
2500
3000
3500


Informational
Count
Informational
0
5000
10000
15000
20000
25000
30000
0
5
10
15
20
25

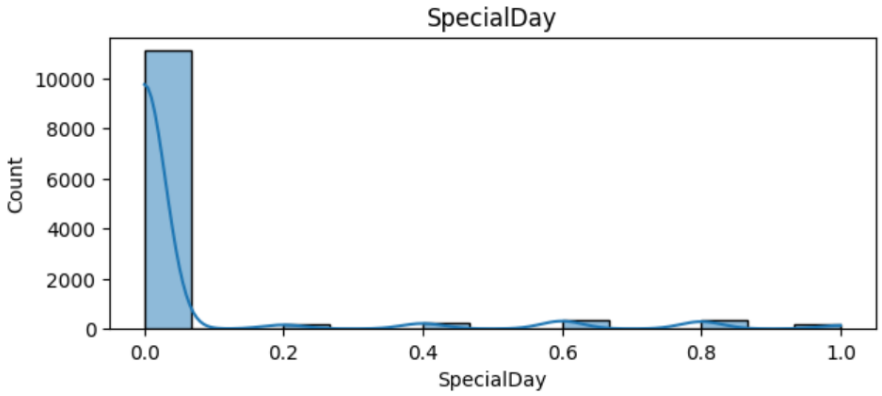

Informational_Duration
Count
Informational_Duration
0
5000
10000
15000
20000
25000
30000
0
500
1000
1500
2000
2500

- Administrative y Administrative_Duration, Information y Information_Duration: Las 4 variables presentan gran cantidad de valores cero, lo que indicaría que son variables poco relevantes sin embargo se analizará la relación con variable a predecir para confirmar esta inferencia.

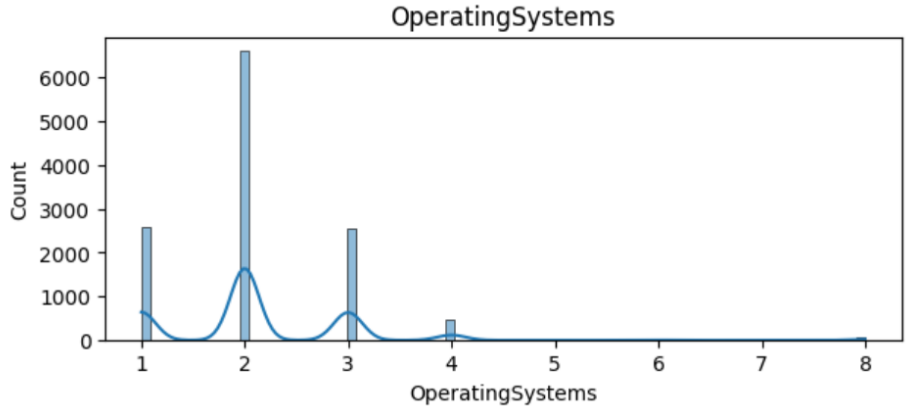


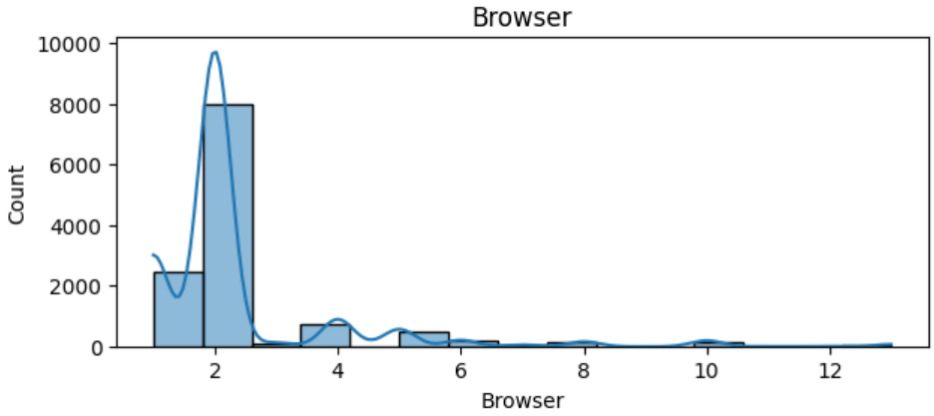


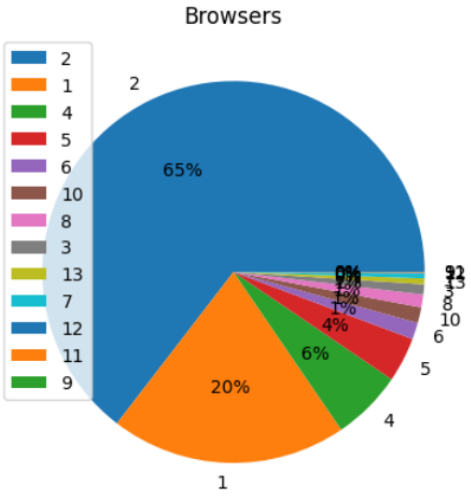
- ProductRelated y Duration: se observan outliers cuando los potenciales clientes ven muchos productos relacionados y pasan mucho tiempo, sin embargo se espera que puedan ser datos importantes y muy relacionados a la compra final.

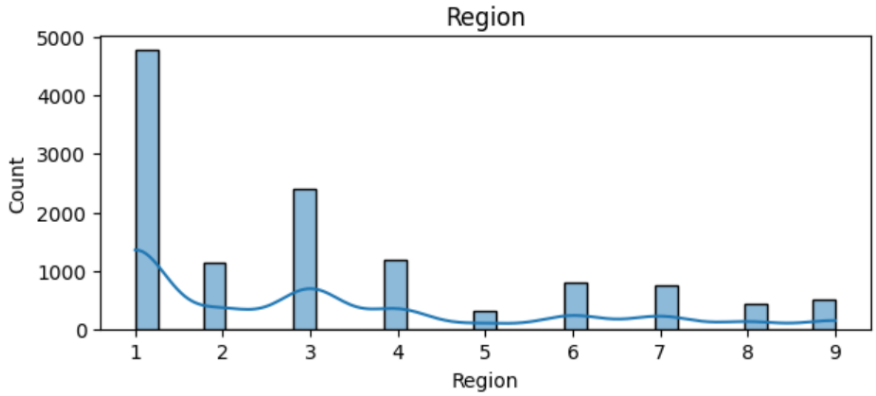


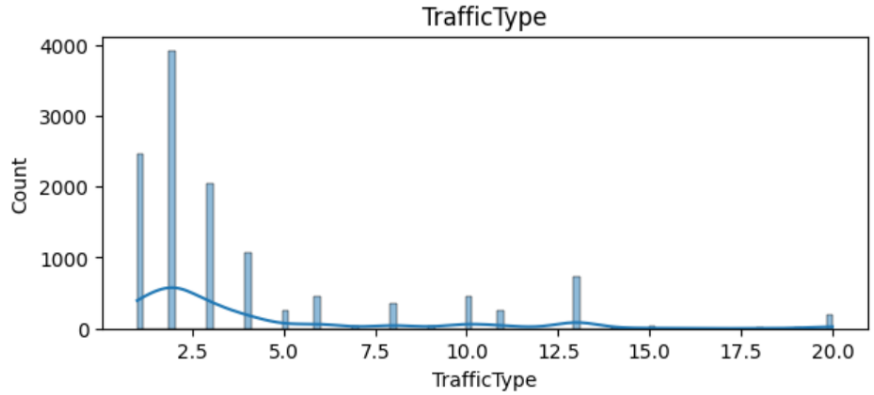
- SpecialDay: ocurre lo mismo, los valores mayores a 0 se presentan como outliers pero debemos ver si tienen mucha relación con la compra final





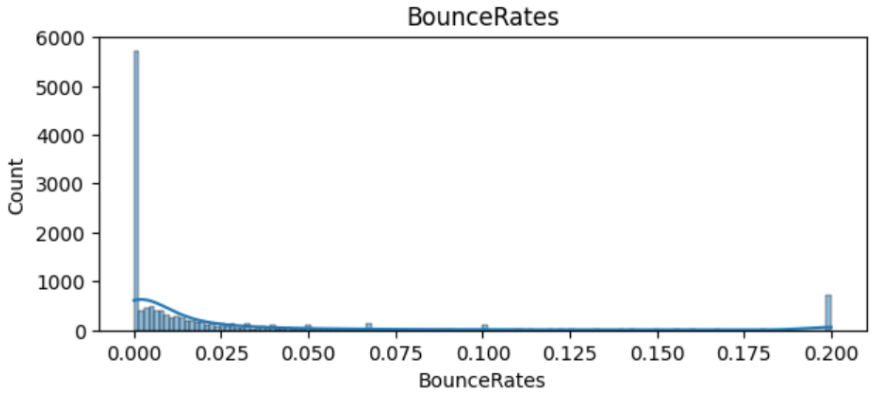


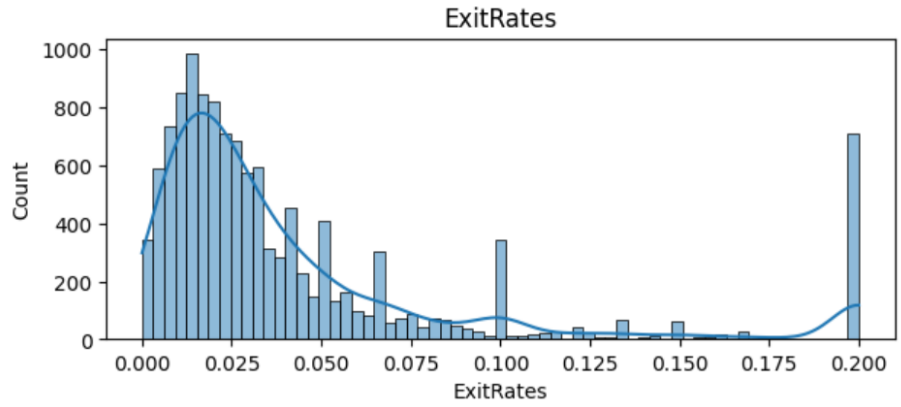


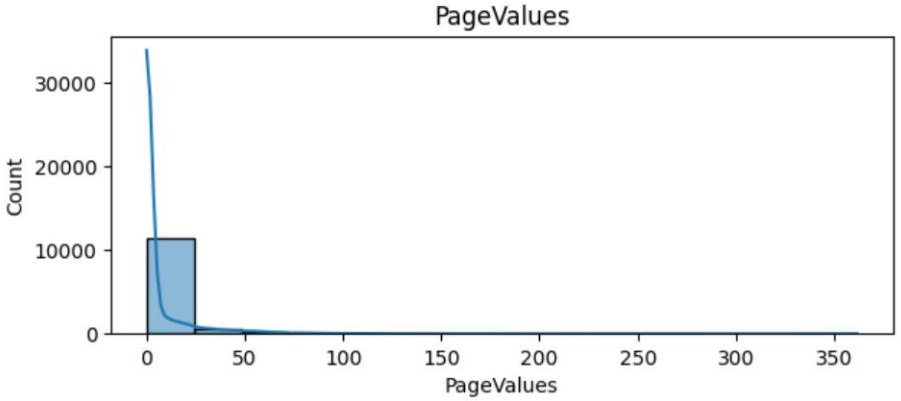


* Region, Browser, Traffic Type y OperatingSystems: se observan casos donde no se realizan muchas transacciones. Se recomienda realizar algún tratamiento de Outliers.

Los navegadores más utilizados son el 2, el 1 y el 4. Entre esos 3 navegadores se realizan el 90% de las visitas a la página.

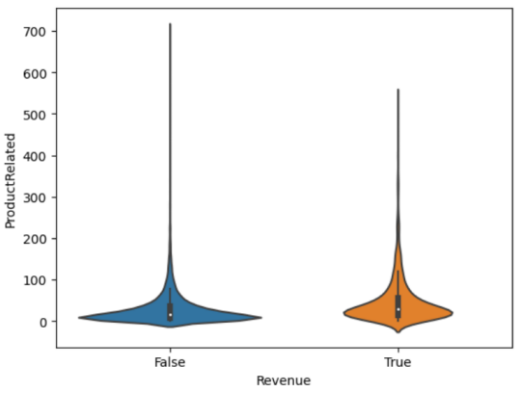


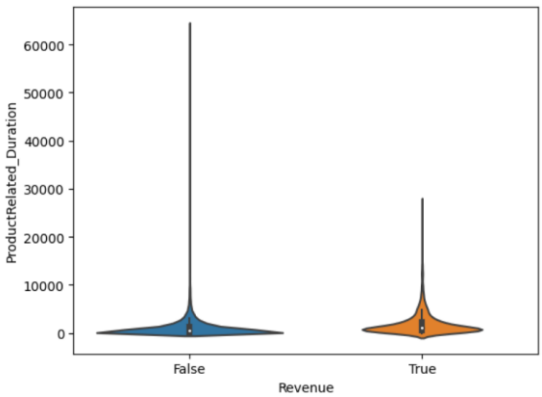


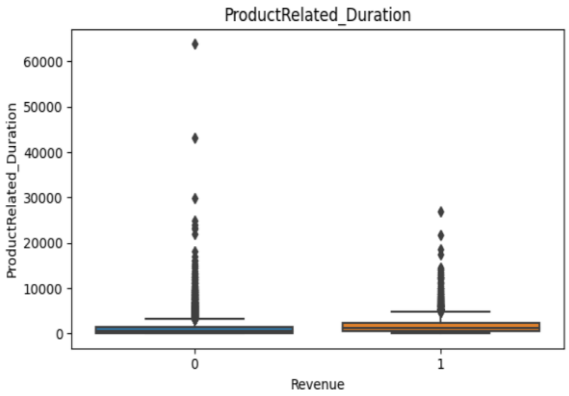


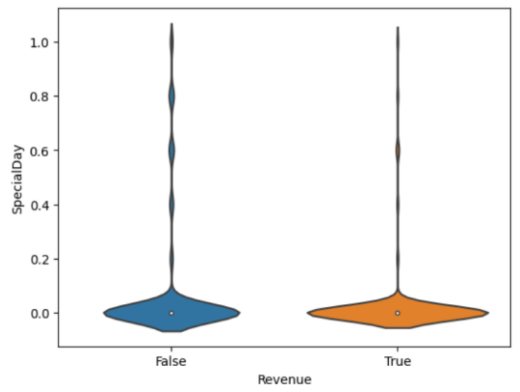
Estas últimas 3 variables no presentan observaciones relevantes en el análisis univariado, se analizará su relación respecto a la variable a predecir.

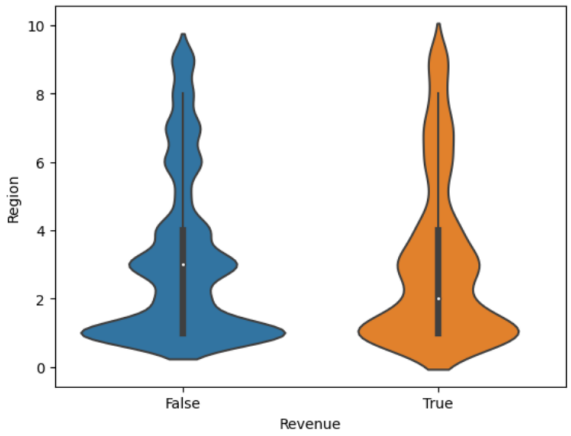
Análisis de variables sobre la variable a predecir

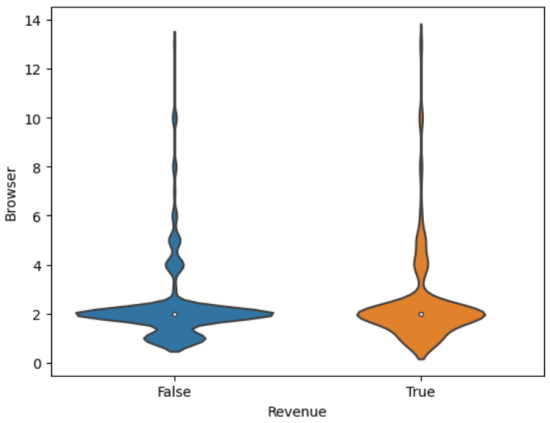


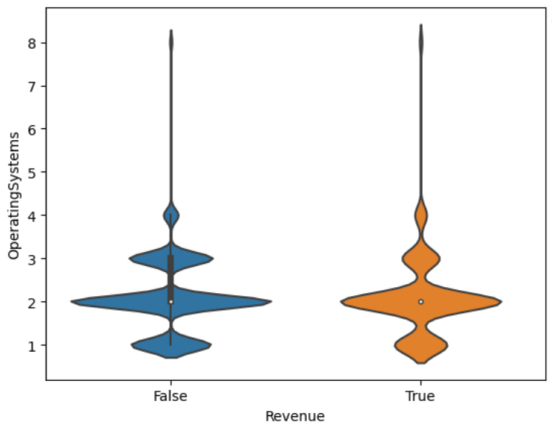


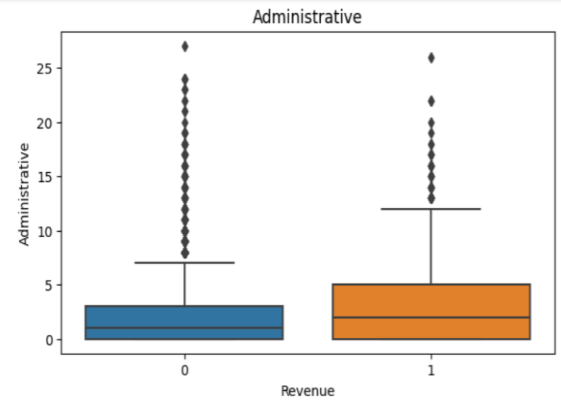

ProductRelated_Duration
ProductRelated_Duration
60000
50000
40000
30000
20000
10000
0
0
1
Revenue

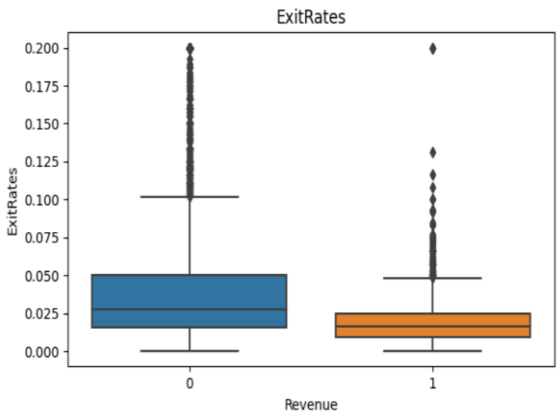

SpecialDay
1.0
0.8
0.6
0.4
0.2
0.0
False
True
Revenue


Region
10
8
6
4
2
0
False
True
Revenue

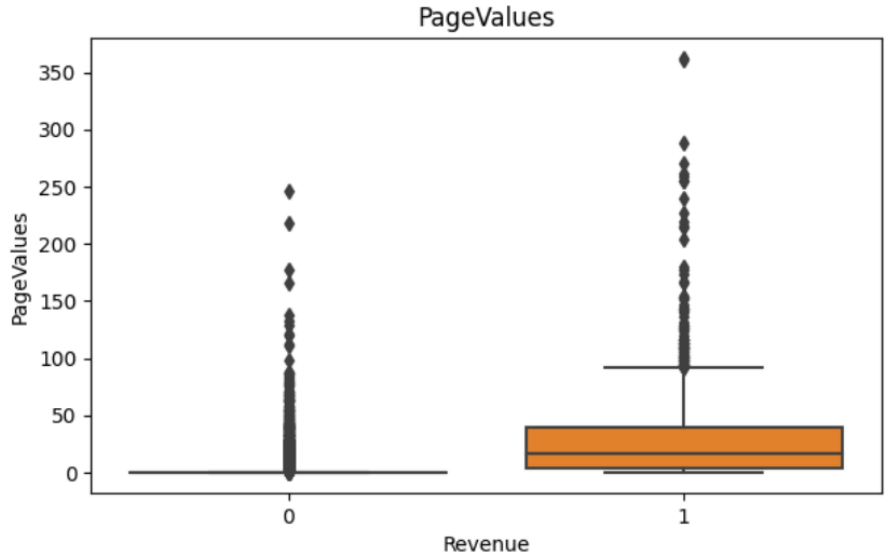

Browser
14
12
10
8
6
4
2
0
False
True
Revenue


OperatingSystems
8
7
6
5
4
3
2
1
False
True
Revenue


Administrative
Administrative
25
20
15
10
5
0
0
1
Revenue


ExitRates
ExitRates
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
1
Revenue

Se observan Outliers en la variable ProductRelated_Duration y se decide reemplazarlos por el tope intercuartil para mejorar el análisis.



La variable PageValues parece ser la que presenta mayor relación directa con la variable a predecir

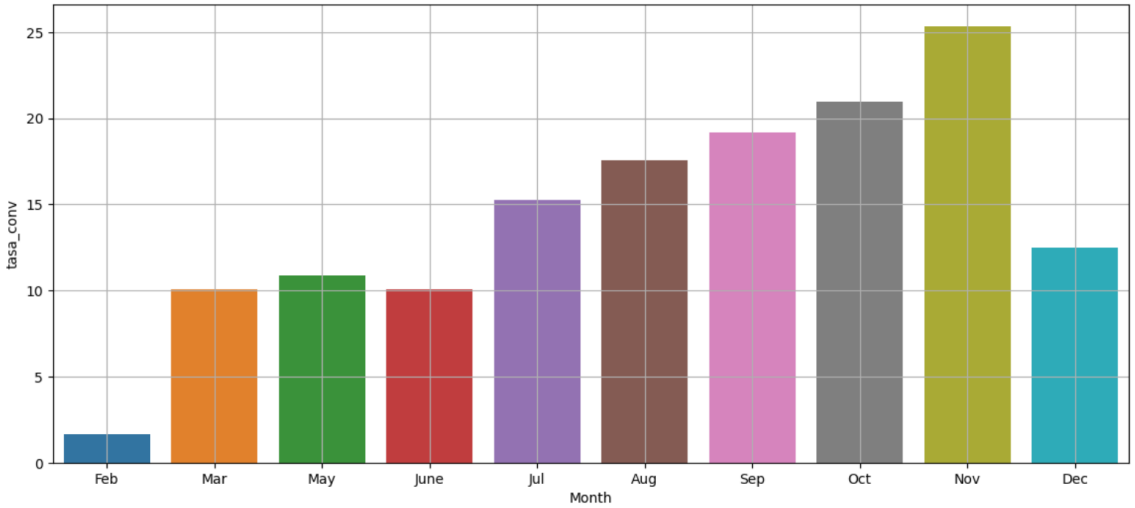
Observando los gráficos precedentes se considera que las principales variables que podrían tener relación sobre la decisión de compra final son las siguientes:

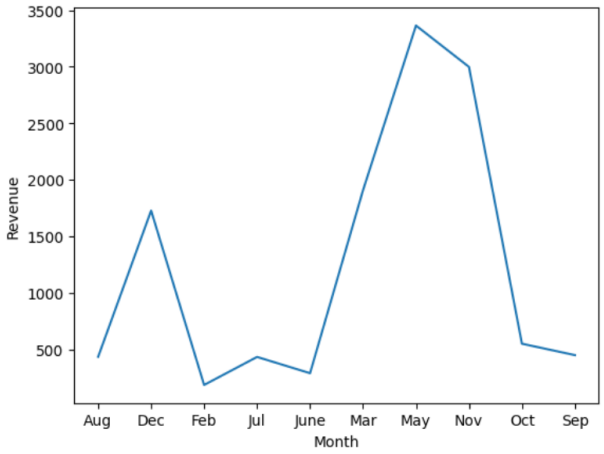
1) Month
2) VisitorType
3) Weekend
4) ProductRelated
5) PageValues
6) Region

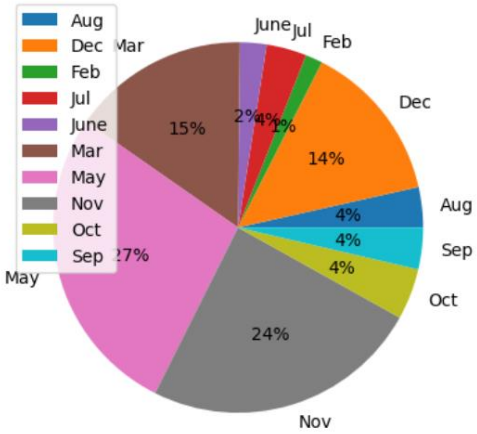
Se decide realizar un análisis de la Tasa de Conversión de estas variables sobre la variable a predecir para confirmar, o no, la relación que tienen con la decisión de compra final.

Análisis de Tasas de Conversión

1) Month



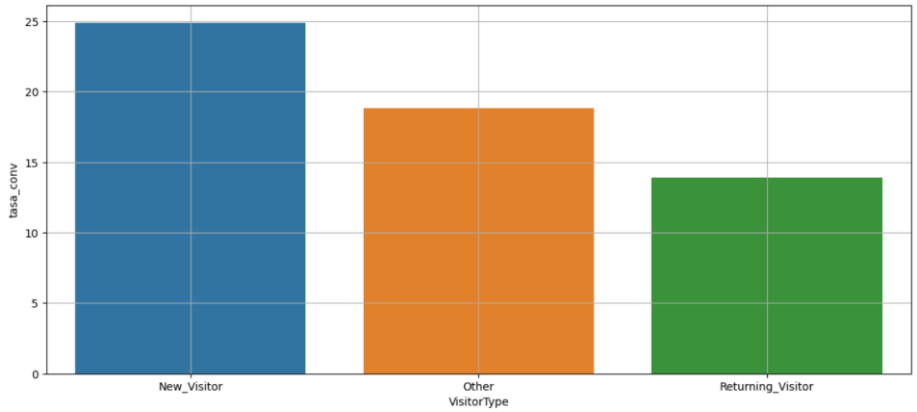




Se observa que los meses que más se utiliza la página son May, Nov, Mar y Dec. En esos 4 meses se acumula el 80% de las visitas.

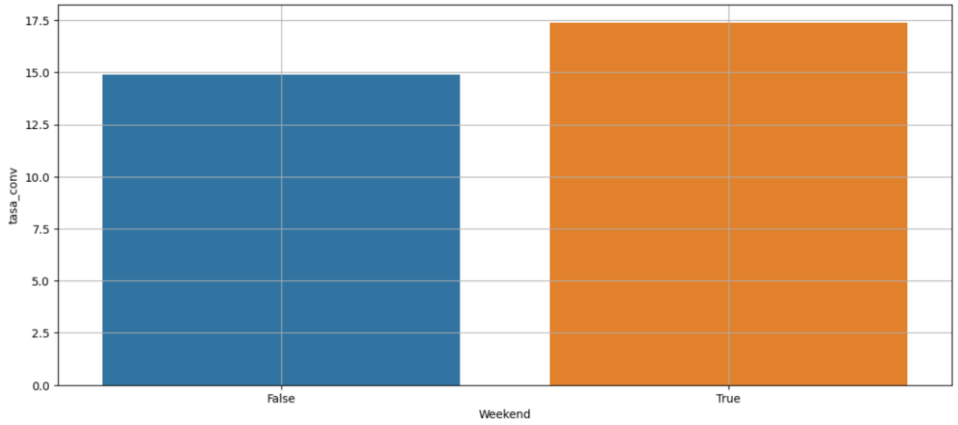
Sin embargo, al observar la tasa de conversión deducimos que Jul, Aug, Sep, Oct y Nov, que son los meses de mayor tasa. Entre el 15% y 26% de las veces que se ingresa a la web se termina comprando el producto, son los meses con mayor decisión de compra.

2) VisitorType

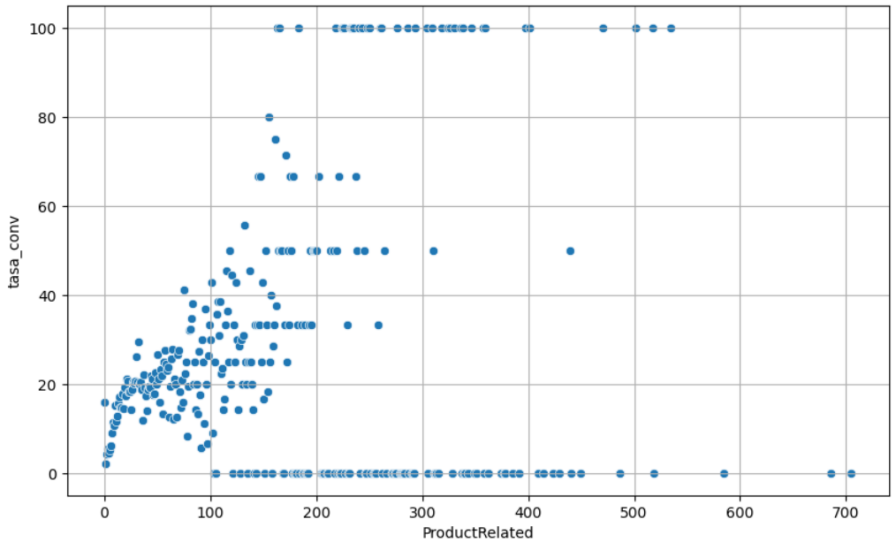


Contrario a lo esperado, los Returning Visitors son los de menor tasa de conversión, es decir, que sólo un 14% de los que vuelven terminan comprando.

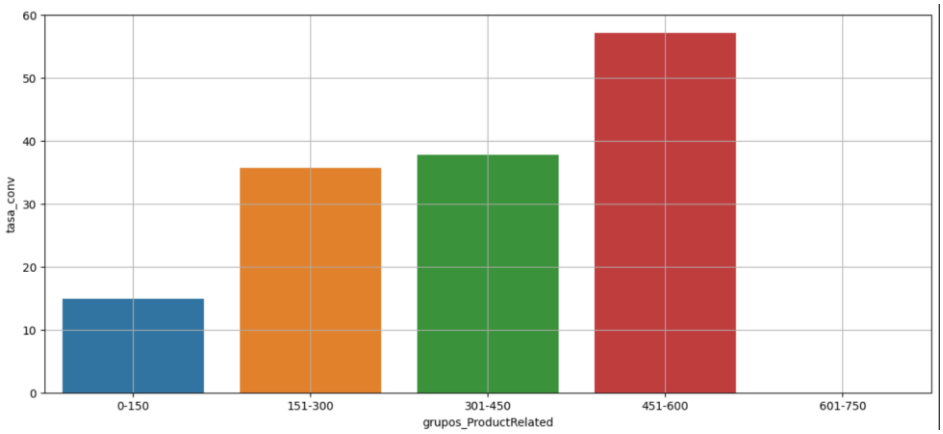
3) Weekend



4) ProductRelated

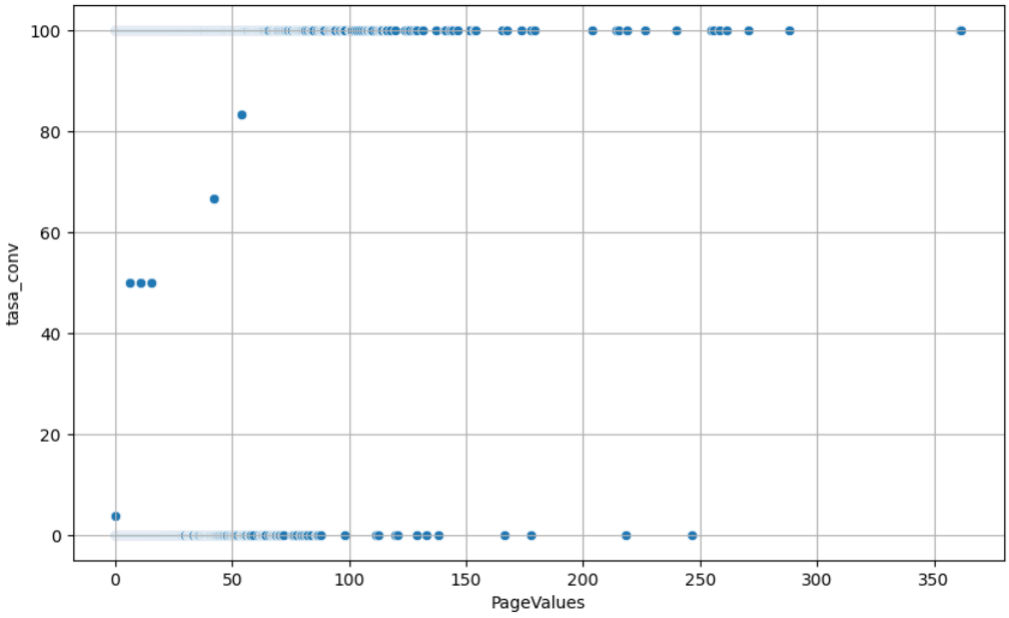


En esta variable se decide conformar grupos a fin de permitir mejorar la visualización y la posibilidad de sacar conclusiones

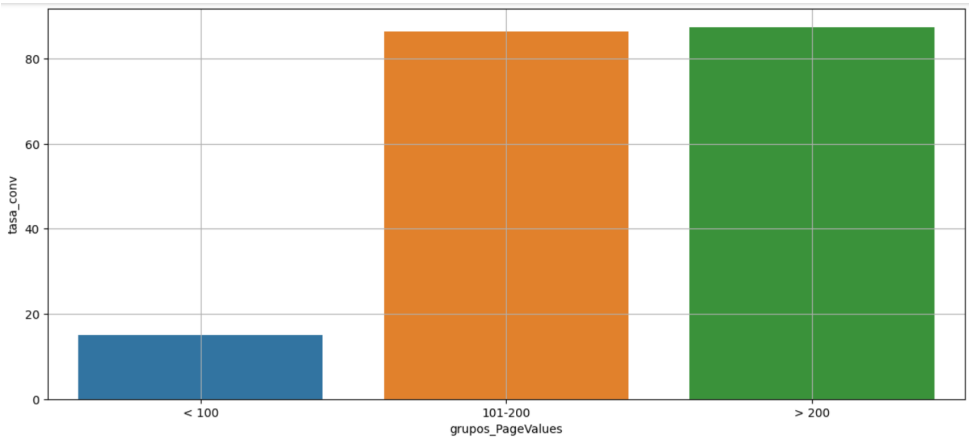


La agrupación propuesta muestran que en el grupo entre 451-600 páginas visitadas de productos relacionados posee una alta tasa de compra final.

5) PageValues

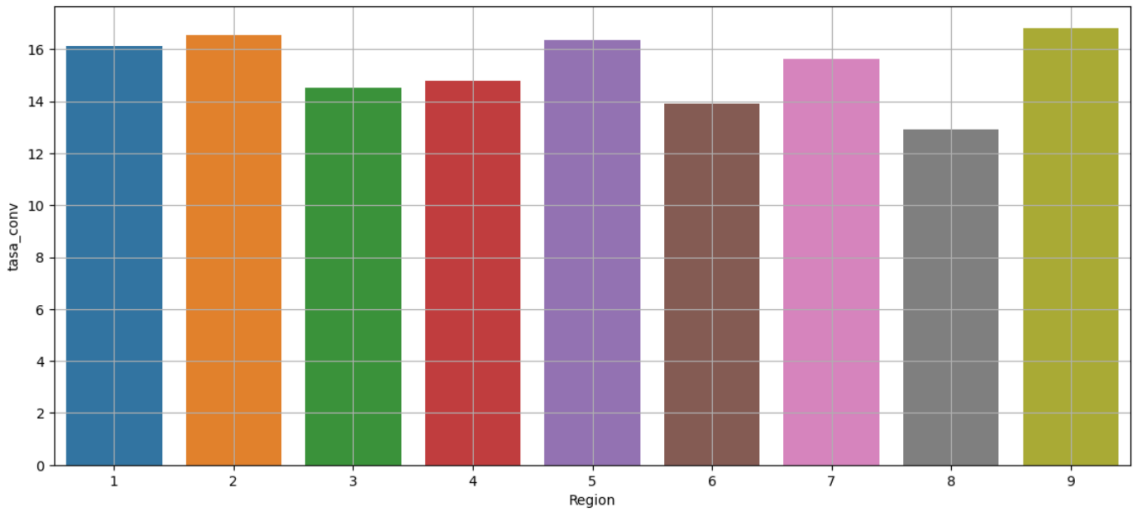


En este caso también se decide conformar grupos para obtener mejores conclusiones



Se confirma lo mencionado anteriormente respecto a la relación de esta variable con la decisión de compra final, para los grupos mayores a 100 un 85% de las veces se termina comprando.

6) Region



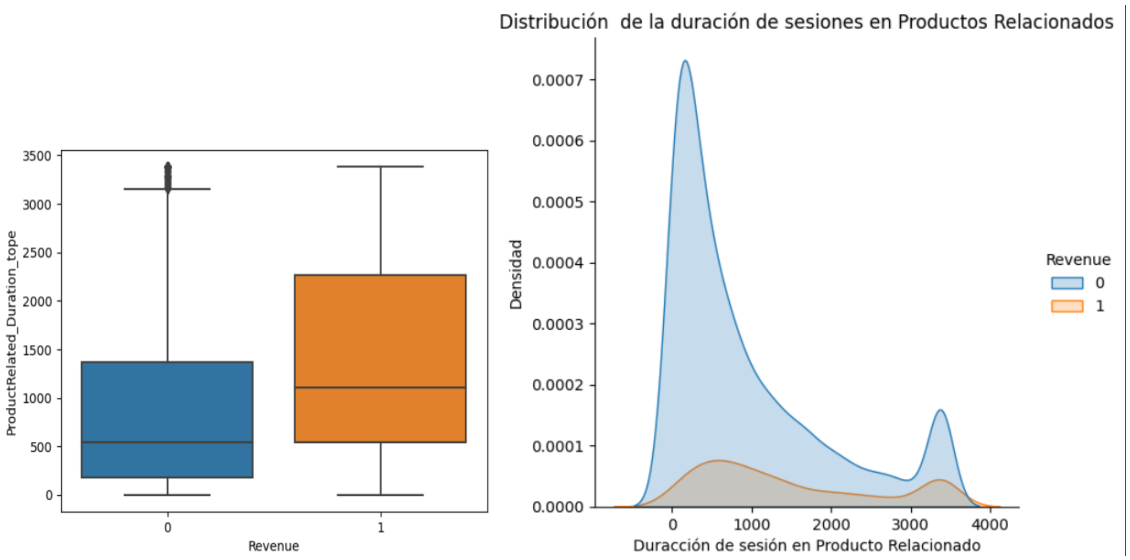
No se observa diferencia por Region

Análisis sobre Hipótesis planteadas:

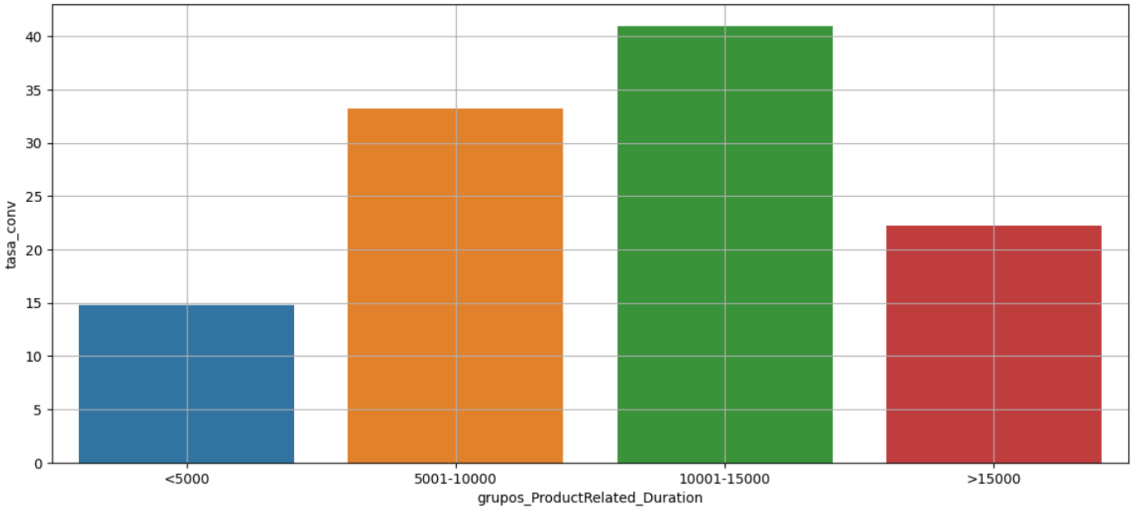
Hipótesis 1: A mayor tiempo de duración en páginas similares mayor probabilidad de venta

Se considera la variable Product Duration como importante ya que el tiempo que las personas dedican a una página se entiende se relaciona en forma directa con el interés de esa persona en el producto.

Se analiza la variable aplicando el tope intercuartil para atenuar el efecto de los Outliers



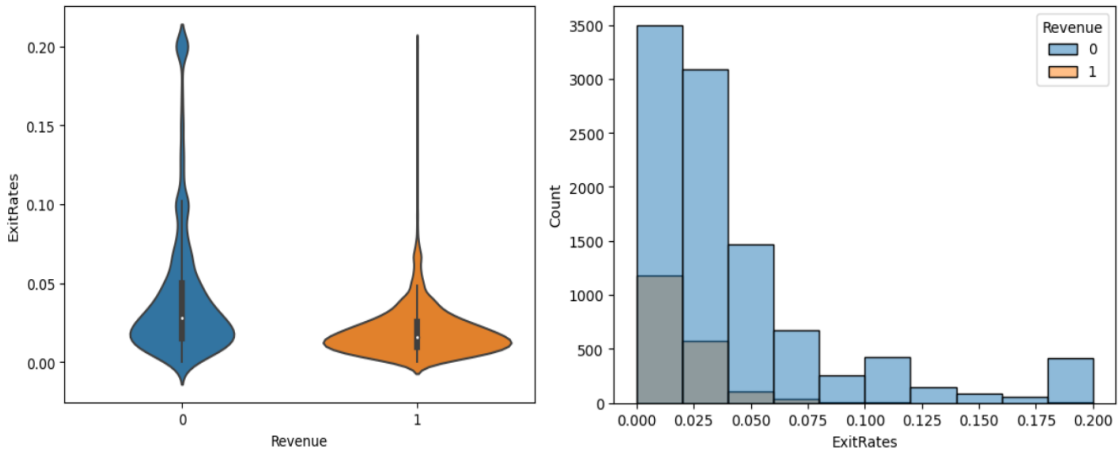
Se establecen grupos para calcular la tasa de conversión de la variable.



Conclusión:

Se observa una tendencia a que la venta se concrete cuando la duración de la sesión en un producto relacionado es mayor, lo cual hace sentido considerando que la gente suele dedicar mayor tiempo a las cosas sobre las que tiene interés. Observando el gráfico de tasa de conversión también se destaca que cuando supera determinado tiempo la decisión de compra disminuye, lo que podría indicar que son usuarios que realizan mucha investigación antes de comprar y terminaron comprando en otro sitio, tal vez por menor precio u otro beneficio. Se recomiendo realizar un análisis de la competencia sobre estos casos.

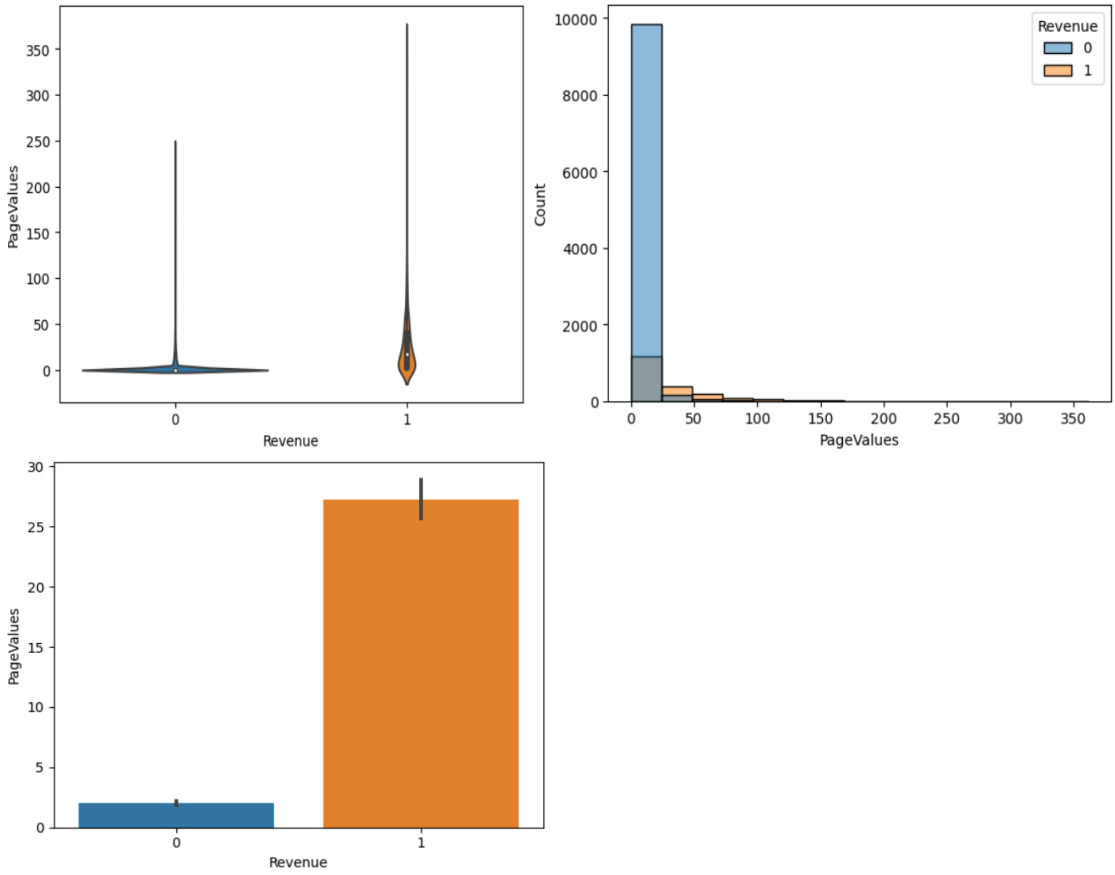
Hipótesis 2: Mayor ExitRate menor probabilidad de compra



Conclusión:

A mayor ExitRate no se observa una diferencia en la decisión de compra, sin embargo podemos concluir que las compras se realizan cuando el ExitRate es bajo.

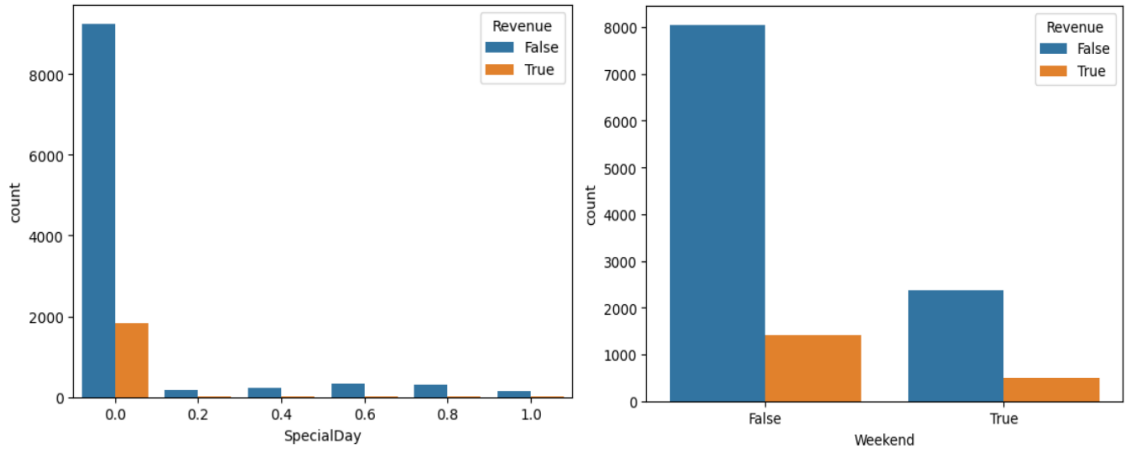
Hipótesis 3: Mayor PageValue mayor probabilidad de compra



Conclusión:

Se observa mayor intención de compra cuando el PageValue es mayor a cero

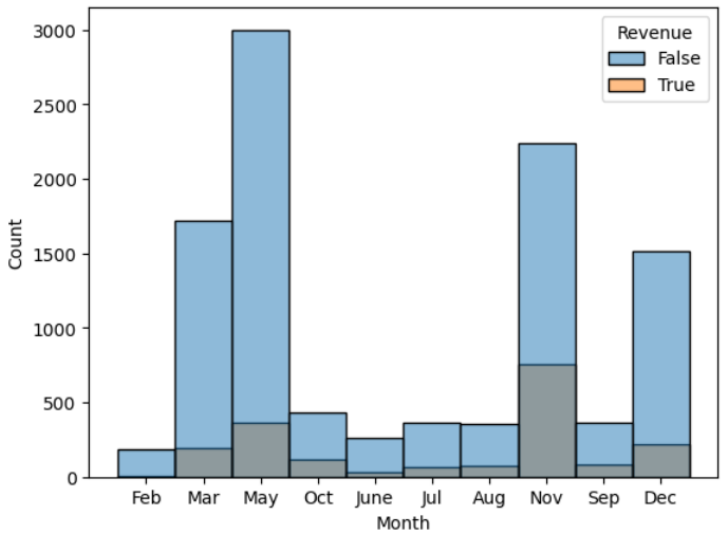
Hipótesis 4: Las ventas crecen los fines de semana y los períodos anteriores a días especiales



Conclusión:

No se observa diferencia en la intención de compra cerca de los fines de semana o días especiales.

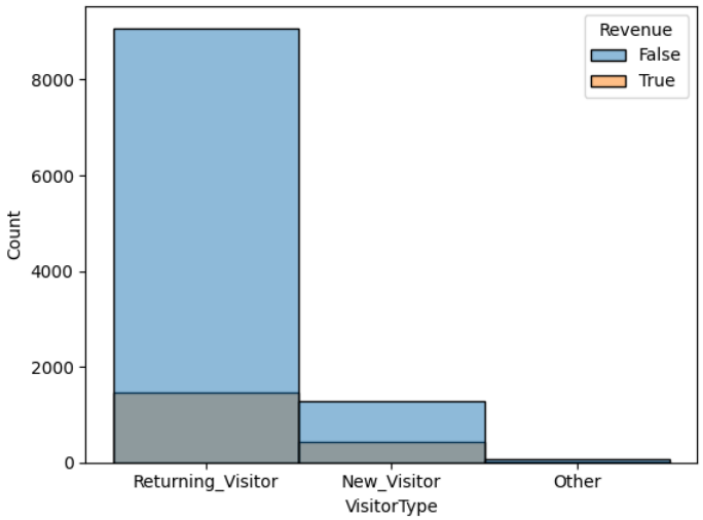
Hipótesis 5: Hay estacionalidad en las compras



Conclusión:

Se observa estacionalidad con una clara mayor intención de compra en el mes de Noviembre.

Hipótesis 6: La mayoría de las compras las realizan al volver a visitar la página



Conclusión:

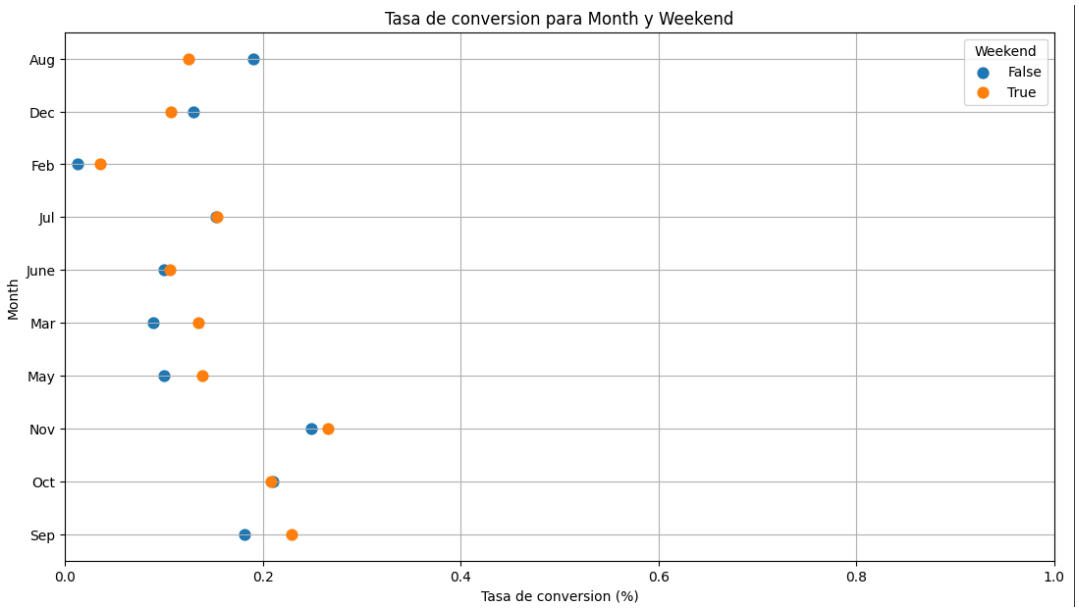
La mayoría de las compras las realizan quienes vuelven a visitar la página, sin embargo también la mayoría de quienes visitan la página son Returning Visitors, el porcentaje de gente que termina comprando no es alto, recordemos que la tasa de compra final era mayor en visitantes nuevos. Se recomienda realizar campañas para incrementar visitantes nuevos.

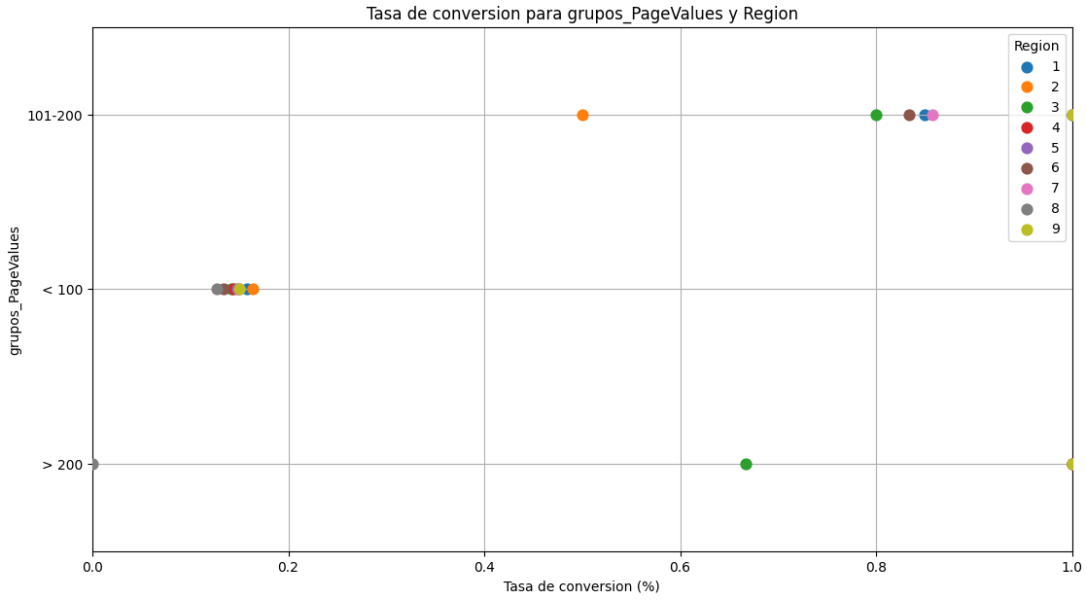
Análisis Bivariado sobre variable a predecir

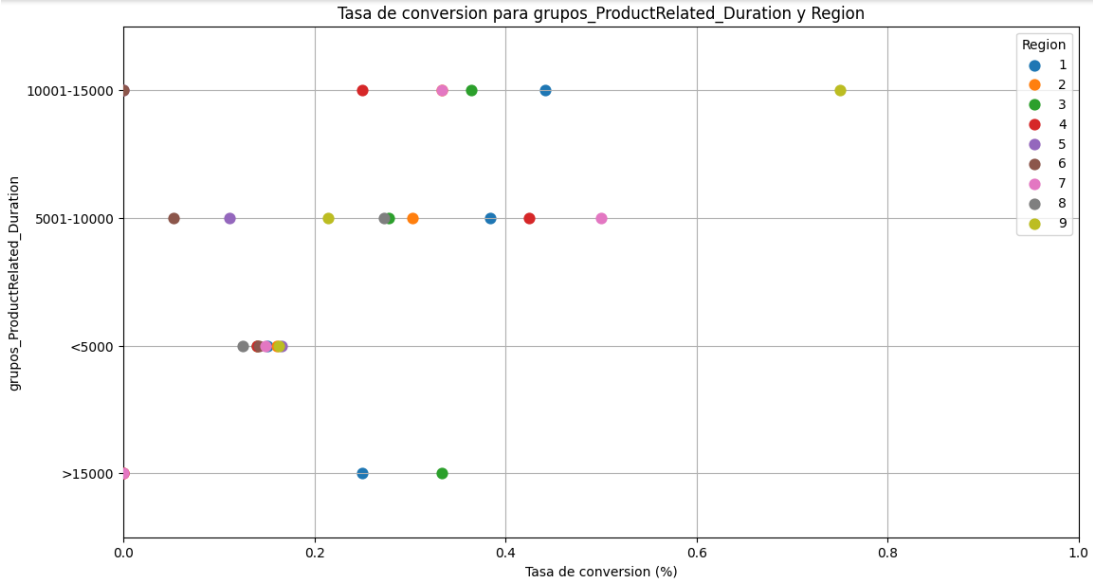
Tasas de conversión Bivariada

Se analizarán las siguientes combinaciones de variables, que se espera tengan correlación:

I. Month y Weekend
II. PageValues y Region
III. Region y ProductRelated


Tasa de conversion para Month y Weekend
Weekend
False
True
Aug
Dec
Feb
Jul
June
Mar
May
Nov
Oct
Sep
Month
0.0
0.2
0.4
0.6
0.8
1.0
Tasa de conversion (%)


Tasa de conversion para grupos_PageValues y Region
Region
1
2
3
4
5
6
7
8
9
101-200
< 100
> 200
grupos_PageValues
0.0
0.2
0.4
0.6
0.8
1.0
Tasa de conversion (%)


Tasa de conversion para grupos_ProductRelated_Duration y Region
Region
1
2
3
4
5
6
7
8
9
10001-15000
5001-10000
<5000
>15000
grupos_ProductRelated_Duration
0.0
0.2
0.4
0.6
0.8
1.0
Tasa de conversion (%)

No se observa una correlación significativa entre estas variables y la decisión de compra final.

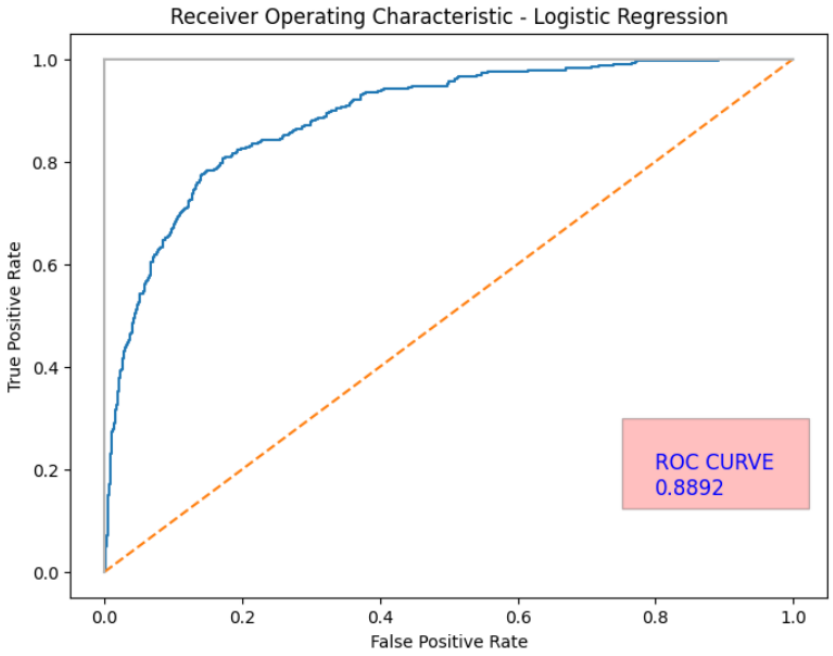
Fase de Modelado

Modelo 1: Regresión Logística

Como primer método se utilizó una Regresión Logística quitando las variables consideradas menos relevantes.

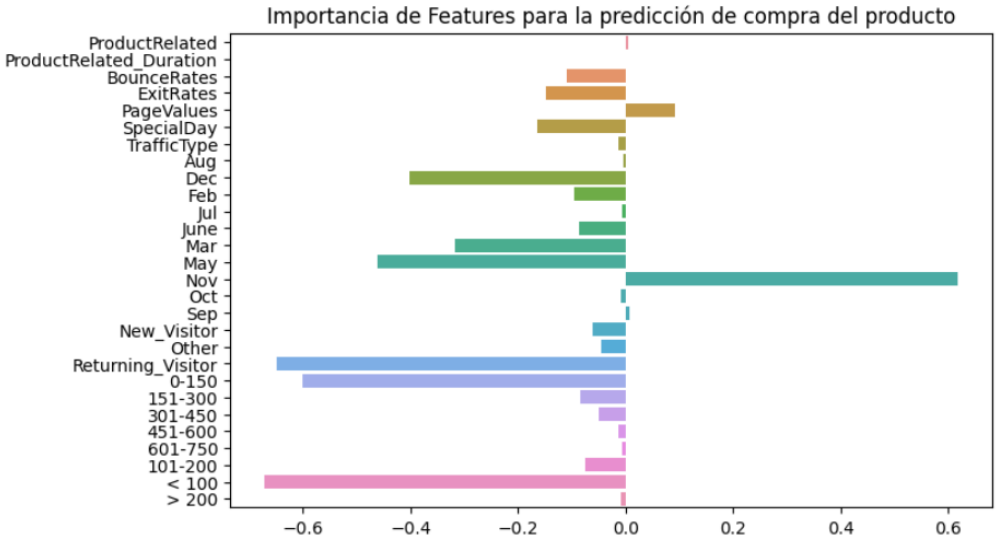
Además se desarrollaron varias técnicas para intentar mejorar el funcionamiento del algoritmo como la codificación de variables categóricas a numéricas utilizando el método get dummies y se integraron al dataset las agrupaciones mencionadas precedentemente para las variables PageValues y ProductRelated.

A partir de este modelo obtuvimos un buen resultado con una Curva ROC cercana al 89%



En el siguiente gráfico podemos observar cuáles resultaron ser las principales variables para la predicción, observándose una importancia notable en las siguientes:

- Noviembre
- Returning_Visitor
- El grupo de 0-150 de la variable ProductRelated
- El grupo < 100 de la variable PageValues

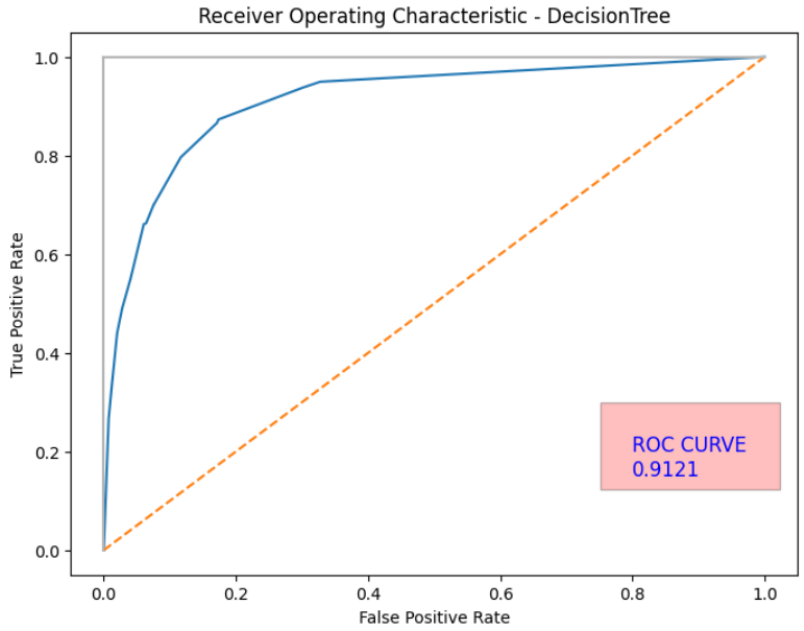


Modelo 2: Árbol de Decisión

Para este modelo se analizará el poder predictivo utilizando todas las variables en un árbol de decisión.

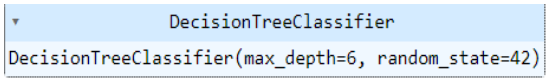
Con este modelo se obtuvo una leve mejora del Accuracy, consiguiendo un 89,72% de precisión.

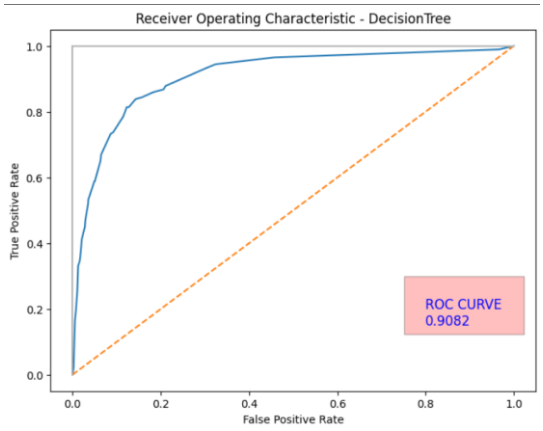
```
DecisionTreeClassifier
DecisionTreeClassifier(max_depth=4, random_state=42)
```

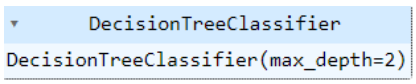


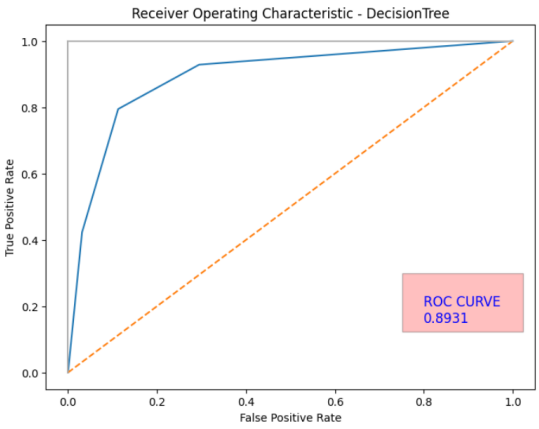
Como podemos observar el modelo mejora levemente, incrementando su AUC en 2 puntos porcentuales

Si modificamos los hiperparámetros del árbol, no se observan diferencias importantes:

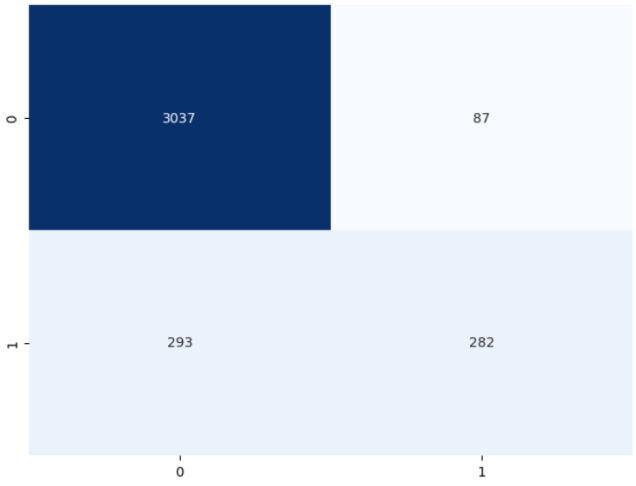








Matriz de confusión:



Observando la matriz de confusión podemos agregar las siguientes conclusiones. El modelo tiene una Precisón del 76.42% que resulta aceptable pero se observa que el modelo tiene mayor capacidad para predecir los Negativos, considerando que su Accuracy es bastante superior, siendo del 89.72%.

Respecto al Recall podemos decir que es bajo, del 49.04% lo que indica que al modelo se le escapan muchos positivos mientras que la Especificidad resulta alta con un 97.21%, es decir, que clasifica muy bien los Negativos.

Sería conveniente buscar un modelo que mejore el porcentaje del Recall a fin de clasificar mejor los positivos y poder realizar campañas sobre esos clientes ofreciéndoles precios o combinaciones de productos para que compren mayor cantidad.

Esta menor capacidad para predecir los clientes que compraron se debe a que el dataset se encuentra desbalanceado.

Modelo 3: PCA

Se analizó un modelo utilizando 4 componentes principales a fin de observar el fucionamiento.

Para obtener un buen modelo utilizando PCA necesitamos de varios componentes, pero de todos modos se aprecia que el PCA no resulta un método efectivo para el set de datos ya que empeora el poder predictivo. El problema probablemente sea generado debido a que algunas variables no son continuas, lo cual afecta negativamente al uso de PCA.

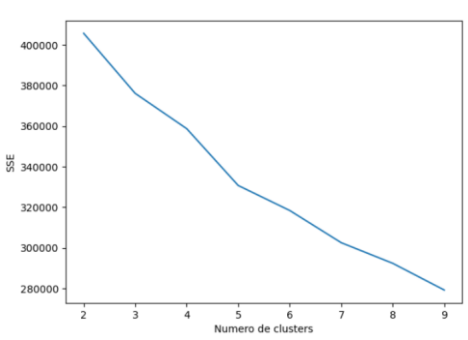
Modelo 4: KNN

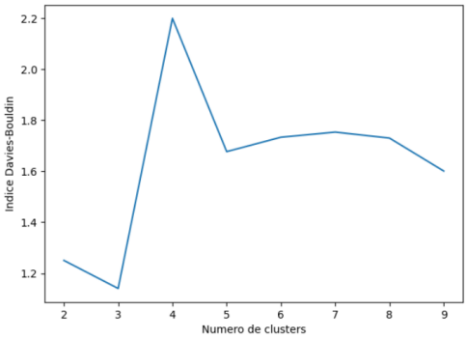
Probamos el modelo KNN definiendo los 3 vecinos más próximos y observamos una sensible caída en la capacidad de predicción del modelo.

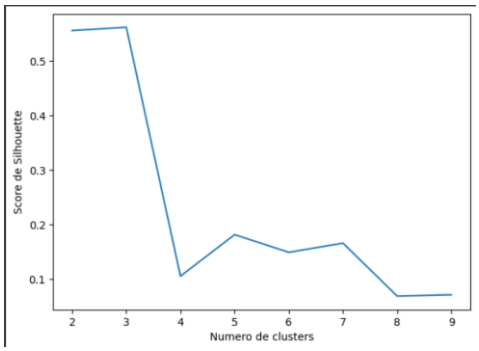
	precision	recall	f1-score	support
False	0.88	0.95	0.91	3124
True	0.51	0.30	0.38	575
accuracy			0.85	3699
macro avg	0.70	0.63	0.65	3699
weighted avg	0.82	0.85	0.83	3699

También se observa un menor Recall, lo cual tampoco permite identificar correctamente ese grupo.

Se procedió a realizar distintos métodos a fin de verificar la cantidad de clusters óptima para el modelo y obtuvimos que





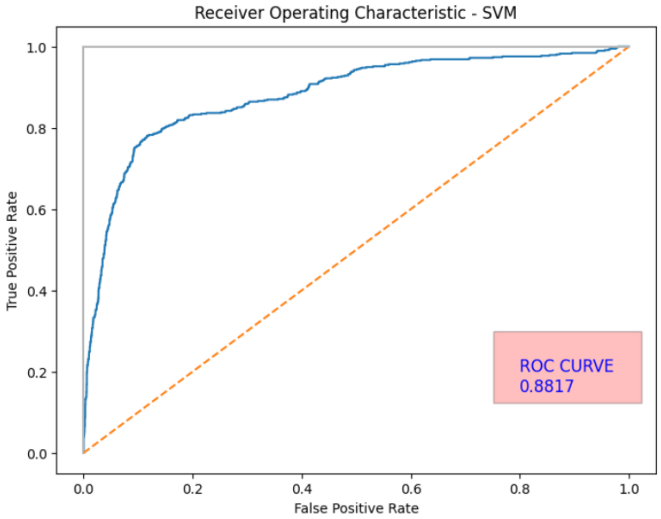


Utilizando los métodos del codo, Davies-Bouldin y Silhouette confirmamos que la cantidad de clusters óptima son 3.

Modelo 5: SVM

Utilizando el modelo Support Vector Machine también logramos resultados menores a los previos.

```
SVC
SVC(C=10, kernel='linear', probability=True, random_state=42)
```



Modelo 6: XGBoost

A fin de analizar este modelo decidimos utilizar combinaciones diferentes de hiperparámetros para verificar cuál resulta ser la más eficiente y precisa.

Resultados:

- El método de Grid Search tuvo un largo período de ejecución de 35 minutos y encontró que el mejor modelo ten+ia un Accuracy del 90,26%.

- El Randomized Search demoró sólo 25 segundos y encontró que su mejor modelo alcanzaba un Accuracy del 90,07%

- El Halving Grid Search, con un preocesamiento de 5 minutos, concluyó que el mejor modelo encontraba un Accuracy del 89,80%.

- El método de Halving Randomized Search, con un tiempo de ejecución de 3 minutos, encontró el mejor modelo con un Accuracy del 89,78%.

Los métodos de validación utilizados encontraron modelos con un Accuracy muy similar y utilizando hiperparámetros distintos. Considerando el tiempo de ejecución y el Accuracy obtenido se considera que el Halving Randomized Search resulta el mejor método de validación para encontrar el modelo a utilizar.

Conclusión

Tras haber desarrollado diversos modelos predictivos observamos que el modelo óptimo resultó ser el obtenido utilizando el Randomized Search, utilizando el algoritmo de XGBoost con los hiperparámetros definidos en el notebook. Este modelo se procesó con bajo costo computacional en sólo 25 segundos y tuvo un Accuracy del 90%.