Metadata

1. Administrative (Integer): This is the number of pages of this type (administrative) that the user visited.
2. Administrative_Duration (Integer): This is the amount of time spent in this category of pages.
3. Informational (Integer): This is the number of pages of this type (informational) that the user visited.
4. Informational_Duration (Integer): This is the amount of time spent in this category of pages.
5. ProductRelated (Integer): This is the number of pages of this type (product related) that the user visited.
6. ProductRelated_Duration (Continuous): This is the amount of time spent in this category of pages.
7. BounceRates (Continuous): The percentage of visitors who enter the website through that page and exit without triggering any additional tasks.
8. ExitRates (Continuous): The percentage of pageviews on the website that end at that specific page.
9. PageValues (Integer): The average value of the page averaged over the value of the target page and/or the completion of an eCommerce transaction.
10. SpecialDay (Integer): This value represents the closeness of the browsing date to special days or holidays (eg Mother's Day or Valentine's day) in which the transaction is more likely to be finalized.
11. Month (Categorical): Contains the month the pageview occurred, in string form.
12. OperatingSystems (Integer): An integer value representing the operating system that the user was on when viewing the page.
13. Browser (Integer): An integer value representing the browser that the user was using to view the page.
14. Region (Integer): An integer value representing which region the user is located in.
15. TrafficType (Integer): An integer value representing what type of traffic the user is categorized into.
16. VisitorType (Categorical): A string representing whether a visitor is New Visitor, Returning Visitor, or Other.
17. Weekend (Binary): A boolean representing whether the session is on a weekend.
18. Revenue (Binary): A boolean representing whether or not the user completed the purchase.

df_clean = df.copy() → Deliver df_clean trước Modeling

Label Encoding for categorical value

1. Month: Jan=1, Feb=2, Mar=3, Apr=4, May=5, Jun=6, Jul=7, Aug=8, Sep=9, Oct=10, Nov=11, Dec=12
2. VisitorType: Other=0, New Visitor=1, Returning Visitor=2
3. Weekend: False=0, True=1
4. Revenue: False=0, True=1

Outlier detection

IQR (Interquartile Range):
Tính Q1 (25%) và Q3 (75%)
IQR = Q3 − Q1
Outlier nếu < Q1 − 1.5×IQR hoặc > Q3 + 1.5×IQR

Outlier handling

1. Browser, TrafficType, Region, OperatingSystems: không xử lý outlier vì là mã phân loại, IQR không có ý nghĩa.
2. BounceRates, ExitRates, SpecialDay: clip vào [0,1] vì đây là biến tỷ lệ/ordinal có miền hợp lệ cố định.
3. Administrative, Informational, ProductRelated, *_Duration, PageValues: clip 99% (P99) để giữ dữ liệu nhưng giảm ảnh hưởng cực trị vì phân phối lệch phải.