Stage 1 Final Project (EDA)

1. Descriptive Statistics

   - tipe data kolom operating system dapat menggunakan tipe data int,
     tipe data kolom month juga dapat menggunakan int. kolom lainnnya sudah sesuai.

   - Terdapat 12.946 baris data, dengan jumlah fitur 18. Dari 18 fitur tersebut, ada 5 fitur yang memiliki nilai null diantaranya:

     1. Administrative `111` null data
     2. Administrative_Duration `633` null data
     3. ProductRelated_Duration `639` null data
     4. BounceRates `74` null data
     5. OperatingSystems `524` null data

     Selain nilai null, juga terdapat 711 data *Duplicated*

   - Untuk fitur numerik (nums) terdapat *outlier* pada masing-masing fiturnya, dan sebaran nilai masing-masing fitur merupakan sebaran *positively skewed*, karena nilai mean yang lebih besar dari nilai median nya.

   - Sedangkan untuk fitur kategorikal (cats), fitur **revenue** dipilih sebagai target. tetapi atribut ini memiliki *imbalances*, dimana nilai *False/Not Buyer* terdapat sebanyak 10.938 data, sehingga perlu untuk disesuaikan ketika proses training.

2. Univariate Analysis

   untuk kolom numerikal berikut ini memiliki distribusi positively skewed dan juga memiliki outlier:

   - 'administrative'
   - 'administrative_duration'
   - 'informational'
   - 'informational_duration'
   - 'productrelated'
   - 'productrelated_duration'
   - 'bouncerates'
   - 'exitrates'
   - 'pagevalues'

   Untuk tahap preprocessing dapat dilakukan, handling outlier dan feature transformation.

   Untuk kolom kategorikal :

   - 'month' : jumlah data didominasi bulan: May, Nov, Mar, Dec
   - 'weekend' : didominasi oleh nilai 'False'
   - 'specialday' : kunjungan situs mayoritas dilakukan saat, jauh dari specialday (hari khusus)
   - 'region' : observasi menunjukan user region 1 mendominasi

- 'operatingsystem' : yang digunakan banyak user 2, 1, 3, 4
- 'browser' : jenis 2 mendominasi data dari 13 jenis browser
- 'traffictype' : jenis traffic yang paling banyak membawa user merupakan traffic 2, 1, 3
- 'visitortype' : kunjungan mayoritas dilakukan oleh returning_visitor
- 'revenue' : sebanyak 84.48% dari kunjungan tidak melakakukan pembelian / tidak menghasilkan pendapatan

Untuk kolom revenue sebagai target perlu dilakukan imbalances handling
kolom visitortype dan month, dapat dilakukan feature encoding agar dapat dilakukan algoritma korelasi

3. Multivariate Analysis

fitur :

- administrative
- informational
- productrelated

memiliki korelasi dengan target

Pagevalues menjadi fitur yang memiliki korelasi sangat relevan dengan target (0.63).

berdasarkan hasil korelasi heatmap yang ditampilkan, terdapat korelasi yang tinggi antara fitur :

- productrelated dengan productrelated_duration (0.88)
- administrative dengan administrative_duration (0.94)

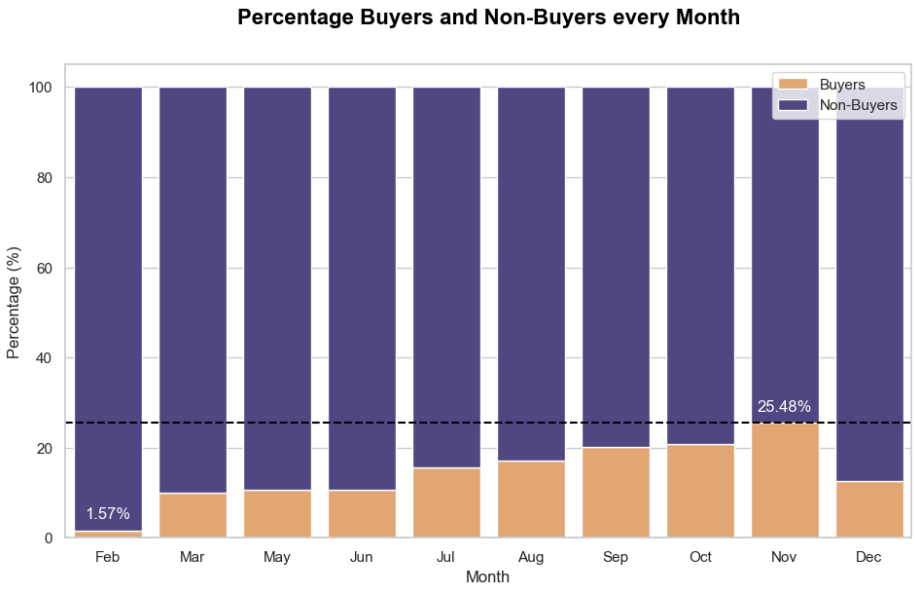
- informational dengan informational_duration (0.95)
- bounce_rates dengan exitrates (0.60)

maka antara salah satu fitur yang berkorelasi tinggi, akan di drop berdasarkan korelasi yang rendah terhadap target **revenue**.

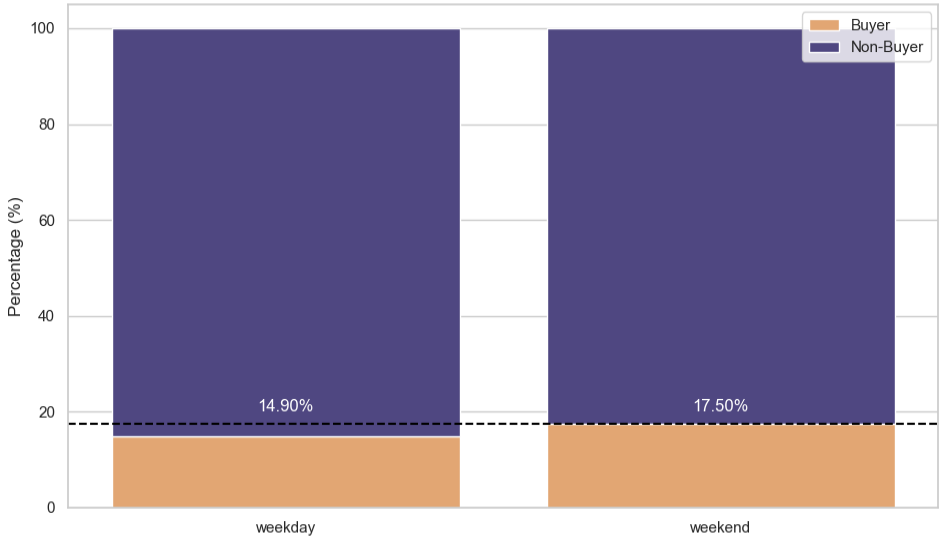
fitur **pagevalues** memiliki korelasi yang tinggi/*relevan* terhadap target. sebesar (0.63)

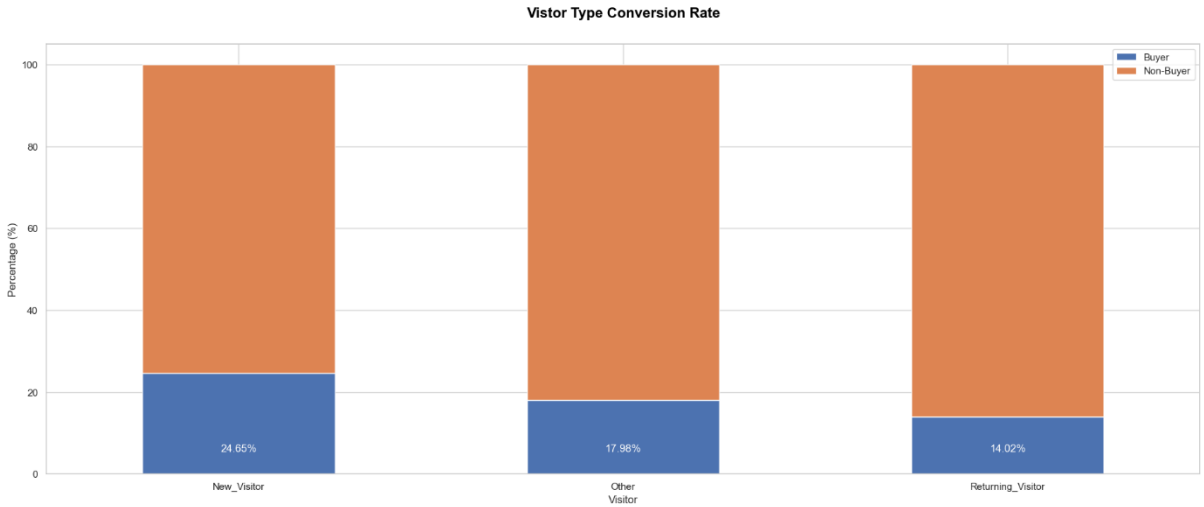
ada kemungkinan fitur month dan visitortype berkorelasi tinggi terhadap target, maka perlu encoding untuk tahap preprocessing dan melihat korelasinya

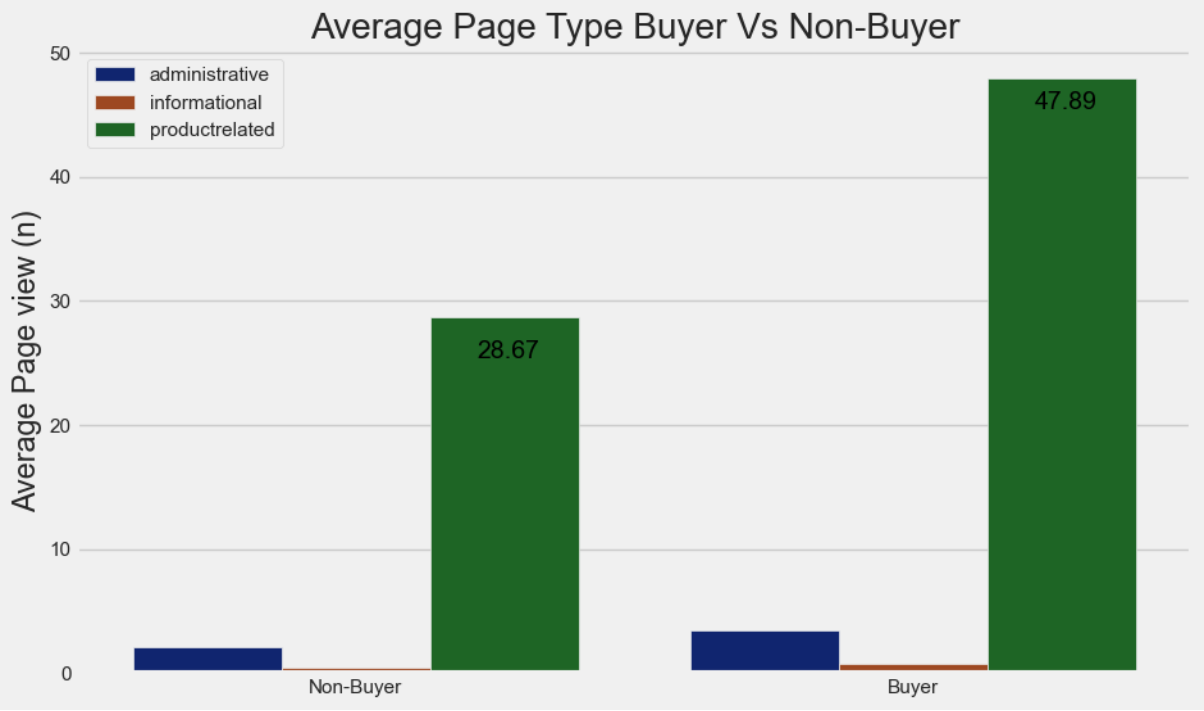
4. Business Insight & Recommendation

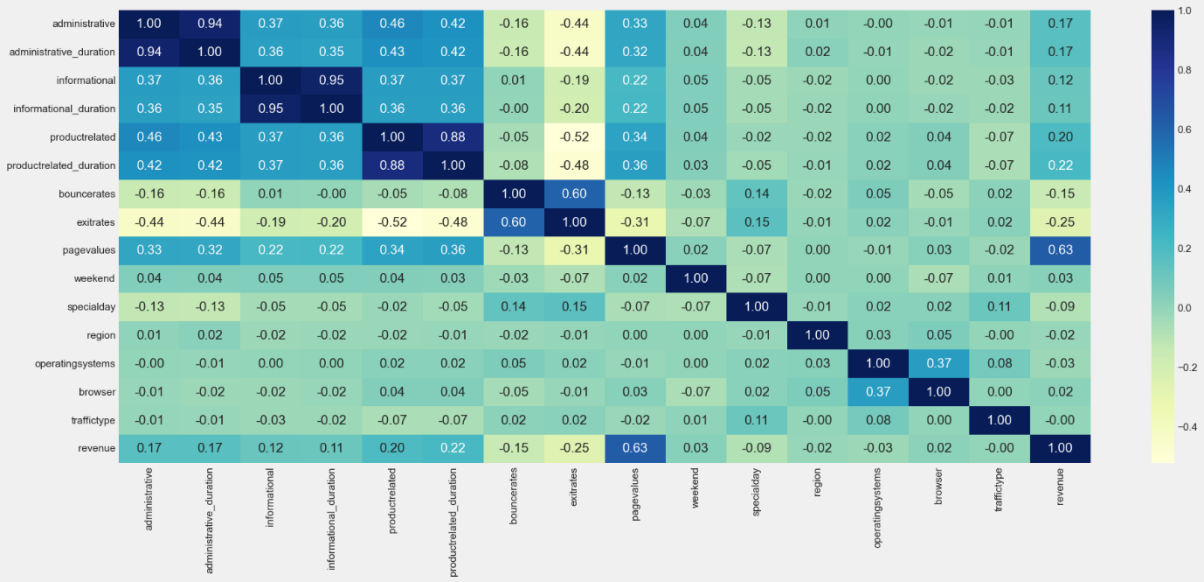


Revenue Rate Weekend / Weekdays






Average Page Type Buyer Vs Non-Buyer
administrative
informational
productrelated
Average Page view (n)
50
40
30
20
10
0
28.67
47.89
Non-Buyer
Buyer


administrative 1.00 0.94 0.37 0.36 0.46 0.42 -0.16 -0.44 0.33 0.04 -0.13 0.01 -0.00 -0.01 -0.01 0.17
administrative_duration 0.94 1.00 0.36 0.35 0.43 0.42 -0.16 -0.44 0.32 0.04 -0.13 0.02 -0.01 -0.02 -0.01 0.17
informational 0.37 0.36 1.00 0.95 0.37 0.37 0.01 -0.19 0.22 0.05 -0.05 -0.02 0.00 -0.02 -0.03 0.12
informational_duration 0.36 0.35 0.95 1.00 0.36 0.36 -0.00 -0.20 0.22 0.05 -0.05 -0.02 0.00 -0.02 -0.02 0.11
productrelated 0.46 0.43 0.37 0.36 1.00 0.88 -0.05 -0.52 0.34 0.04 -0.02 -0.02 0.02 0.04 -0.07 0.20
productrelated_duration 0.42 0.42 0.37 0.36 0.88 1.00 -0.08 -0.48 0.36 0.03 -0.05 -0.01 0.02 0.04 -0.07 0.22
bouncerates -0.16 -0.16 0.01 -0.00 -0.05 -0.08 1.00 0.60 -0.13 -0.03 0.14 -0.02 0.05 -0.05 0.02 -0.15
exitrates -0.44 -0.44 -0.19 -0.20 -0.52 -0.48 0.60 1.00 -0.31 -0.07 0.15 -0.01 0.02 -0.01 0.02 -0.25
pagevalues 0.33 0.32 0.22 0.22 0.34 0.36 -0.13 -0.31 1.00 0.02 -0.07 0.00 -0.01 0.03 -0.02 0.63
weekend 0.04 0.04 0.05 0.05 0.04 0.03 -0.03 -0.07 0.02 1.00 -0.07 0.00 0.00 -0.07 0.01 0.03
specialday -0.13 -0.13 -0.05 -0.05 -0.02 -0.05 0.14 0.15 -0.07 -0.07 1.00 -0.01 0.02 0.02 0.11 -0.09
region 0.01 0.02 -0.02 -0.02 -0.02 -0.01 -0.02 -0.01 0.00 0.00 -0.01 1.00 0.03 0.05 -0.00 -0.02
operatingsystems -0.00 -0.01 0.00 0.00 0.02 0.02 0.05 0.02 -0.01 0.00 0.02 0.03 1.00 0.37 0.08 -0.03
browser -0.01 -0.02 -0.02 -0.02 0.04 0.04 -0.05 -0.01 0.03 -0.07 0.02 0.05 0.37 1.00 0.00 0.02
traffictype -0.01 -0.01 -0.03 -0.02 -0.07 -0.07 0.02 0.02 -0.02 0.01 0.11 -0.00 0.08 0.00 1.00 -0.00
revenue 0.17 0.17 0.12 0.11 0.20 0.22 -0.15 -0.25 0.63 0.03 -0.09 -0.02 -0.03 0.02 -0.00 1.00
administrative
administrative_duration
informational
informational_duration
productrelated
productrelated_duration
bouncerates
exitrates
pagevalues
weekend
specialday
region
operatingsystems
browser
traffictype
revenue
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Insight

- Region 1 memiliki pengunjung paling banyak diantara region lainnya. akan tetapi revenue rate region 2 16.64% menjadi paling tinggi diantara region lainnya.

- Kunjungan user pada platform, yang menghasilkan revenue didominasi pada bulan November 25,48% Revenue Rate, Sementara bulan Februari memiliki kunjungan yang menghasilkan revenue yang paling sedikit 1.57% Revenue Rate (3 buyer).

- Bulan May memiliki kunjungan yang paling banyak diantara yang lain terdapat total kunjungan 3533 akan tetapi, hanya 379 dari total kunjungan yang menghasilkan revenue.

- Kunjungan user pada weekday lebih tinggi dari weekend tetapi revenue rate weekend > weekday **17.5% /14.9%**

- Sesi dilakukan mayoritas oleh Returning Visitors. namun, persentase Buyer pada Returning Visitors secara signifikan lebih sedikit dari Non-Buyers. pada New visitor, proporsi Buyers mendekati proporsi Non-Buyers. hal ini menunjukan bahwa Returning Visitor lebih banyak sesi kunjungannya, tetapi New Visitors mempunyai purchase rate yang lebih tinggi **24.65%.**

- Pengunjung yang memutuskan untuk melakukan pembelian **Buyer**, memiliki nilai rata-rata yang lebih tinggi dari **Non-Buyer**. dalam melihat halaman productrelated **47.89 / 28.67**.

- ketika session melibatkan pagevalues > 0 purchase rate tinggi **56.44%**. Sebaliknya, sesi dengan pagevalues nol menunjukkan purchase rate yang lebih rendah **3.88%**.

Business Recommendation

- untuk region yang masih rendah nilai revenue_rate nya, tim marketing dapat menampilkan halaman web yang memiliki pagavalues > 0, dan juga menampilkan rekomendasi yang relevan dengan halaman web yang yang dikunjungi user (product related). strategi marketing tersebut dapat dilakukan pada weekend, dikarenakan disaat weekend revenue_rate lebih tinggi dibandingkan weekday. maka hal ini dapat membantu meningkatkan revenue platform e-commerce.

Metrics

- Revenue rate