Binary classification of Online Shoppers using multiple features

Binabh Devkota
Roll no: 05
Masters in Data Science

What are we going to discuss?

- Theoritical Aspects
- Approach in this work
- Results
- Limitations
- Conclusion

Before we start

- If you want to follow along with code: binabh.com.np/code.html

Consumer behaviour

- Actions and decisions made by individuals, households, or organizations when they purchase, use, or dispose of products, services, ideas, or experiences
- Influenced by various internal factors, such as personal values, beliefs, attitudes, and motivation and external factors such as marketing messages, brand, social influences, and economic factors and many other may influence (Chovanová et al., 2015)
- Develop products that meet the needs and preferences of their target customers and create effective marketing strategies (Rojhe, 2020)
- It has been an interest of study for a very long time and has exploded in last 50 years (Peighambari et al., 2016)

Consumer behaviour In E-commerce

- Ecommerce has transformed the way consumers shop so we can use consumer behavior data to optimize their online shopping experience and improve customer satisfaction, which can lead to increased sales and customer loyalty(Alshweesh & Bandi, 2022)
- Factors that influence consumer behavior in ecommerce include website design, ease of navigation, product information and reviews, pricing, and payment options and convenience of internet makes customer retention even more challenging(Sv, 2022)

What we have and what we want?

- Dataset derived from https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset.
- The dataset consists of 12,330 customer sessions (rows), 10 numerical and 8 categorical variables (columns).
- We will use ‘Revenue’ (True or False) variable as our dependent variable. The other 17 variables will be our independent variables.

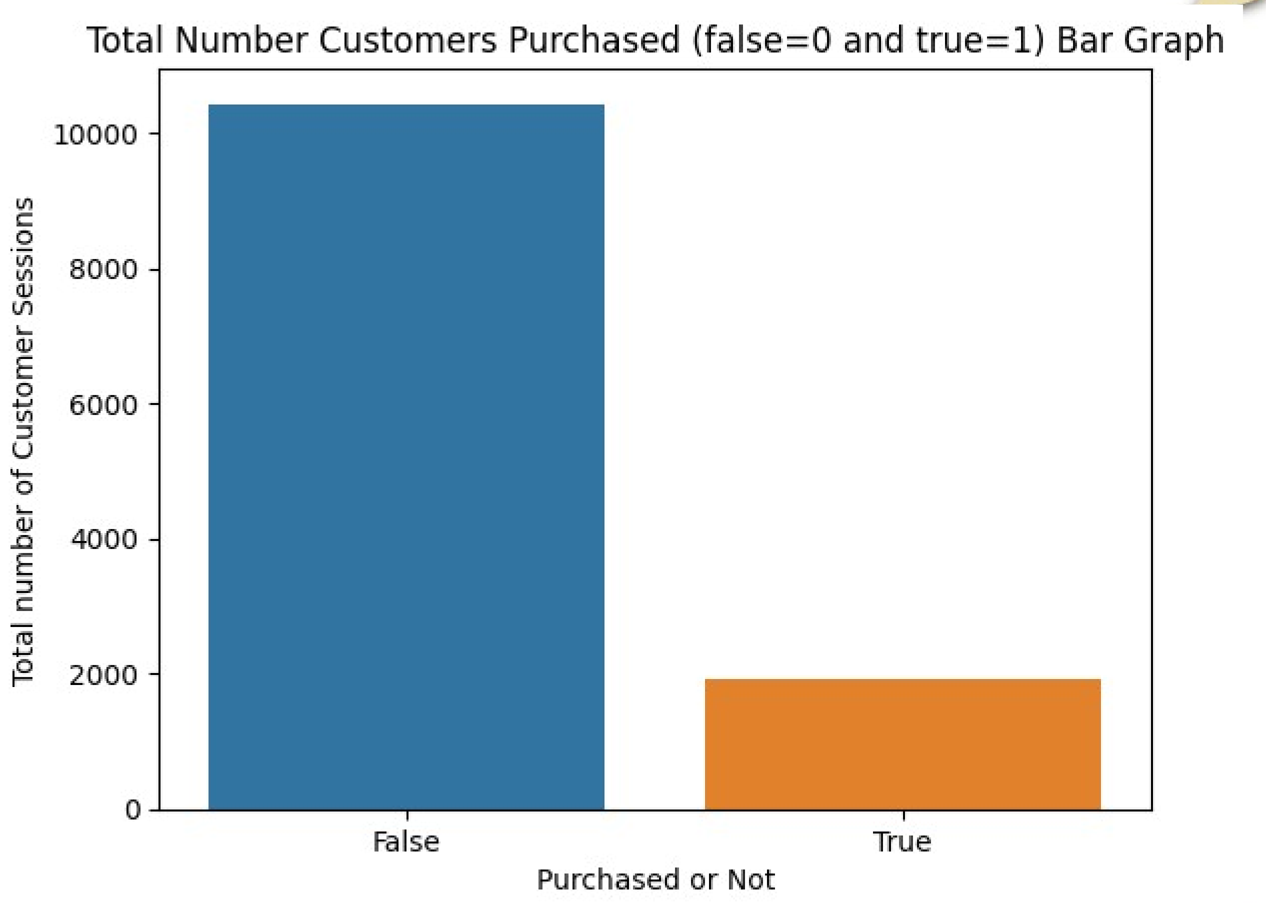
More into the dataset (Numeric Features)

Administrative	Number of pages visited by user about account management
Administrative duration	Time spent by user (in seconds) in account management pages
Informational	Number of times user visits infornationalpages(about us, contact us)
Informational duration	Time spent by user(in seconds) in informational pages
Product related	Number of time use visits product related pages
Product related duration	Time spent by user in product related pages
Bounce rate	Opens one page and leaves
Exit rate	Opens multiple pages and leaves
Page value	Number of pages visited by user
Special day	Closeness of site visiting time to a special day

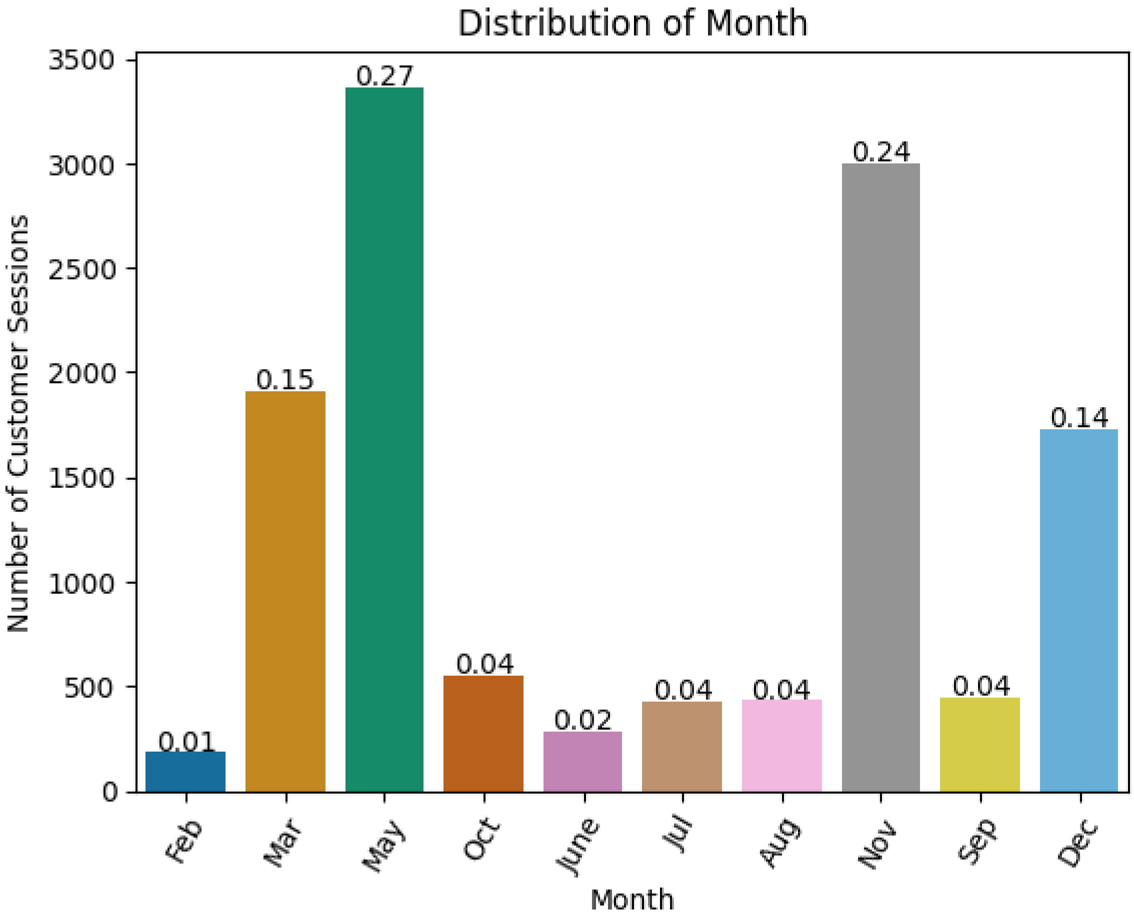
More into the dataset (Categorical Features)

Operating system	Operating system used by user
Browser	Browser used by user
Region	Geographic Region of user
TrafficType	Traffic source (banner ad, sms, direct)
VisitorType	New, returning, other
Weekend	Is visiting date weekend
Month	Month of visit
Revenue	Has visit been finalized with transaction

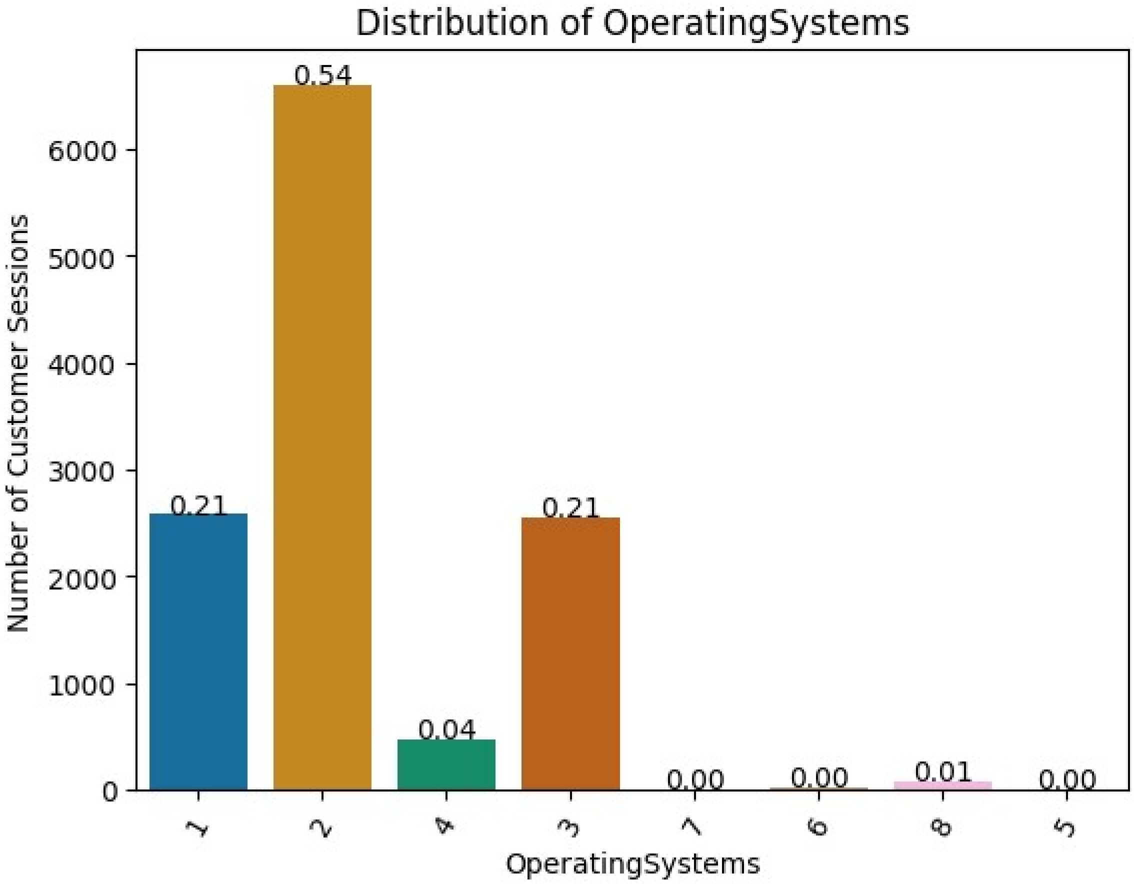
Visualizing categorical features (Revenue)



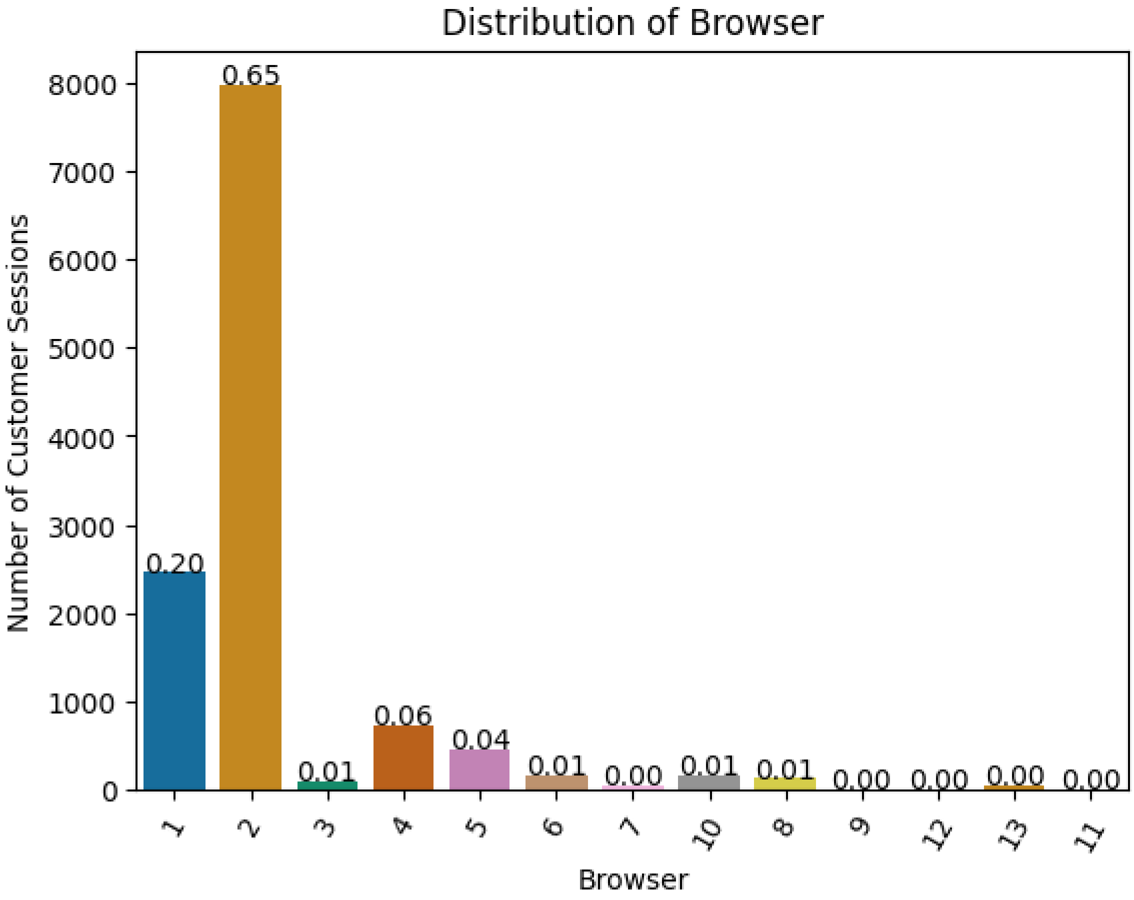
Visualizing categorical features (Month)



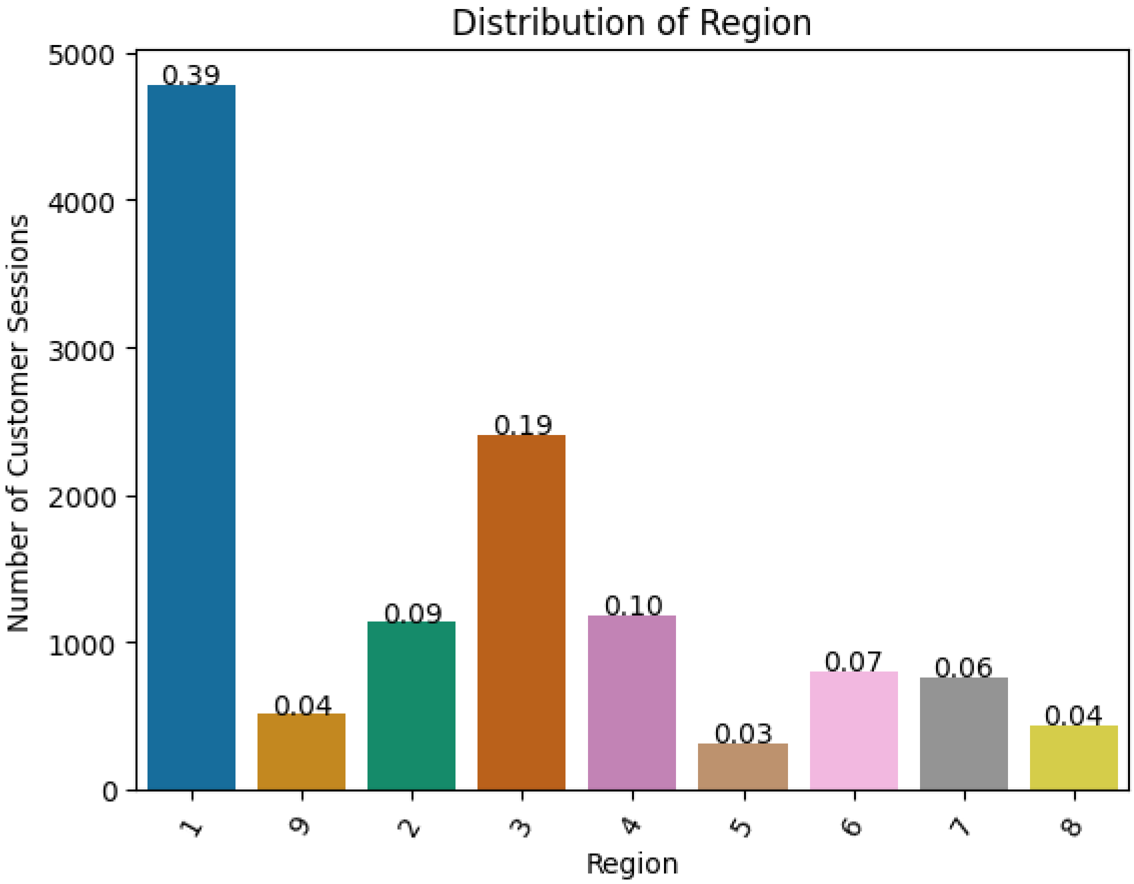
Visualizing categorical features (OS)



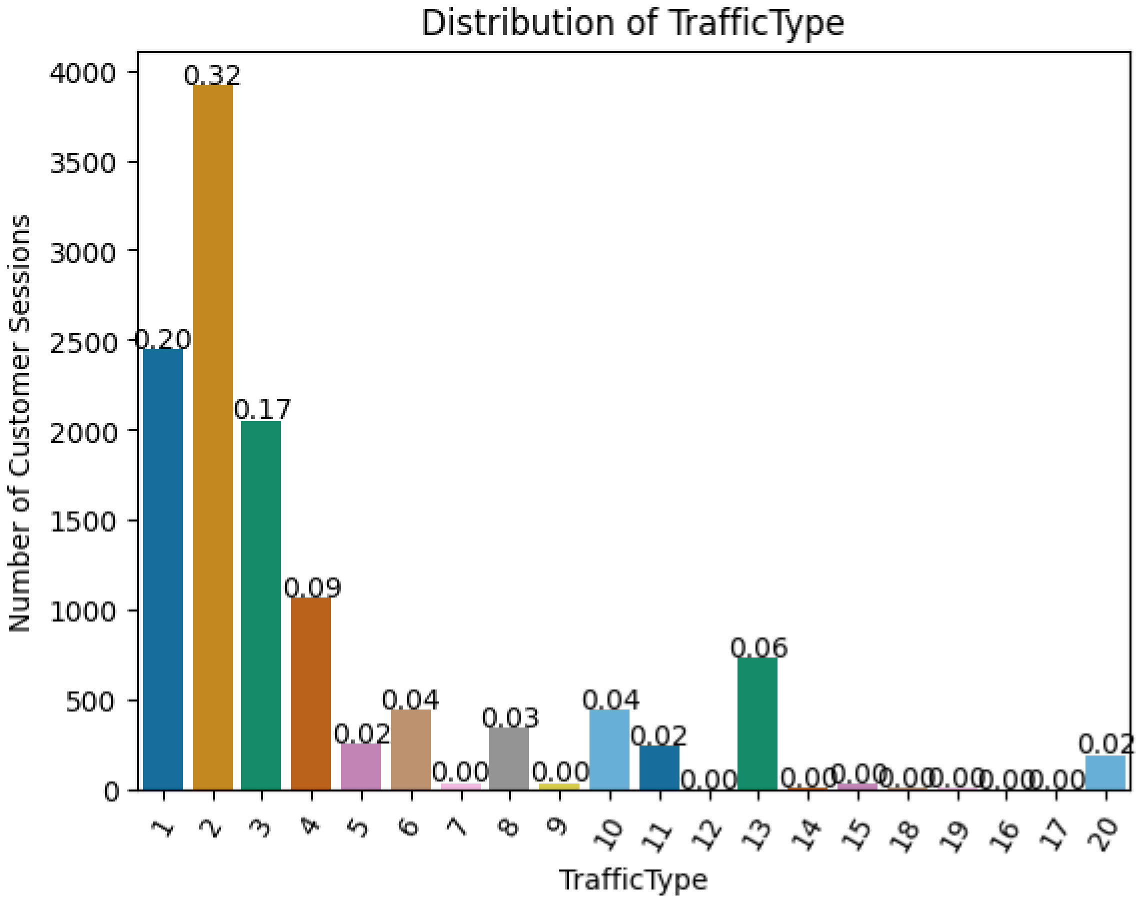
Visualizing categorical features (Browser)



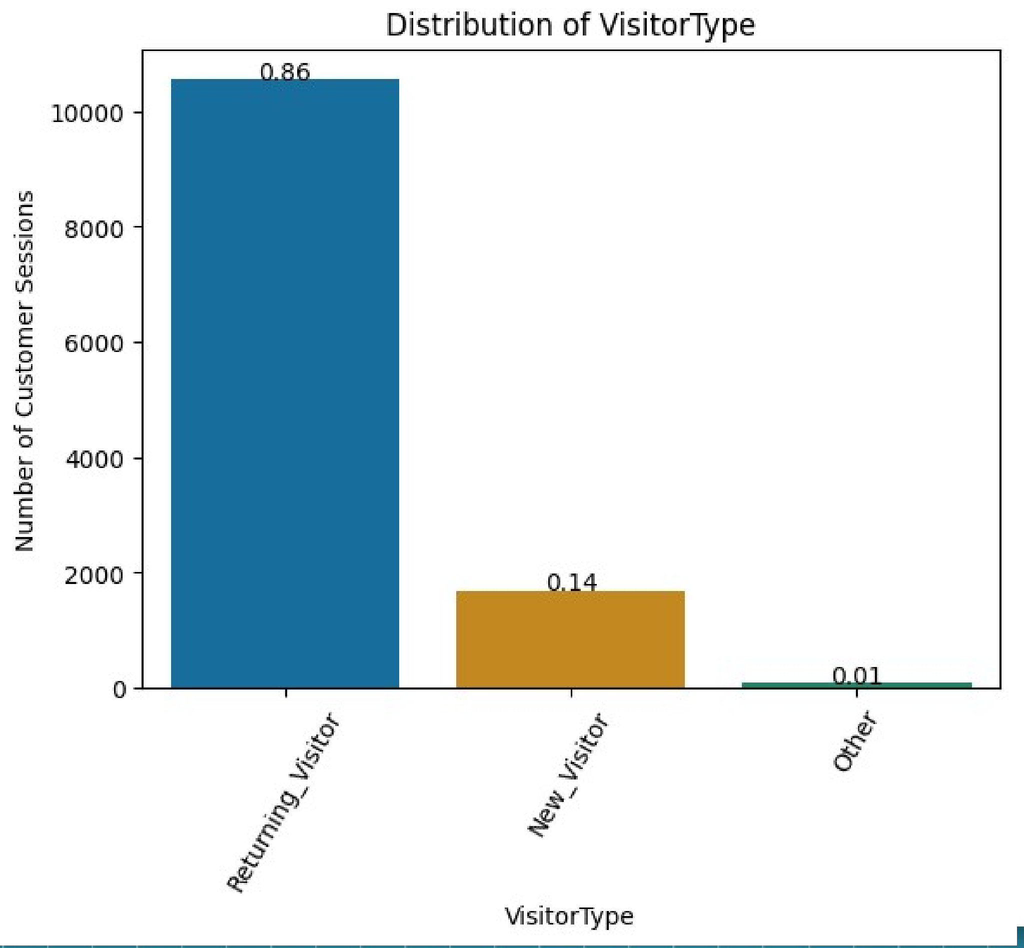
Visualizing categorical features (Region)



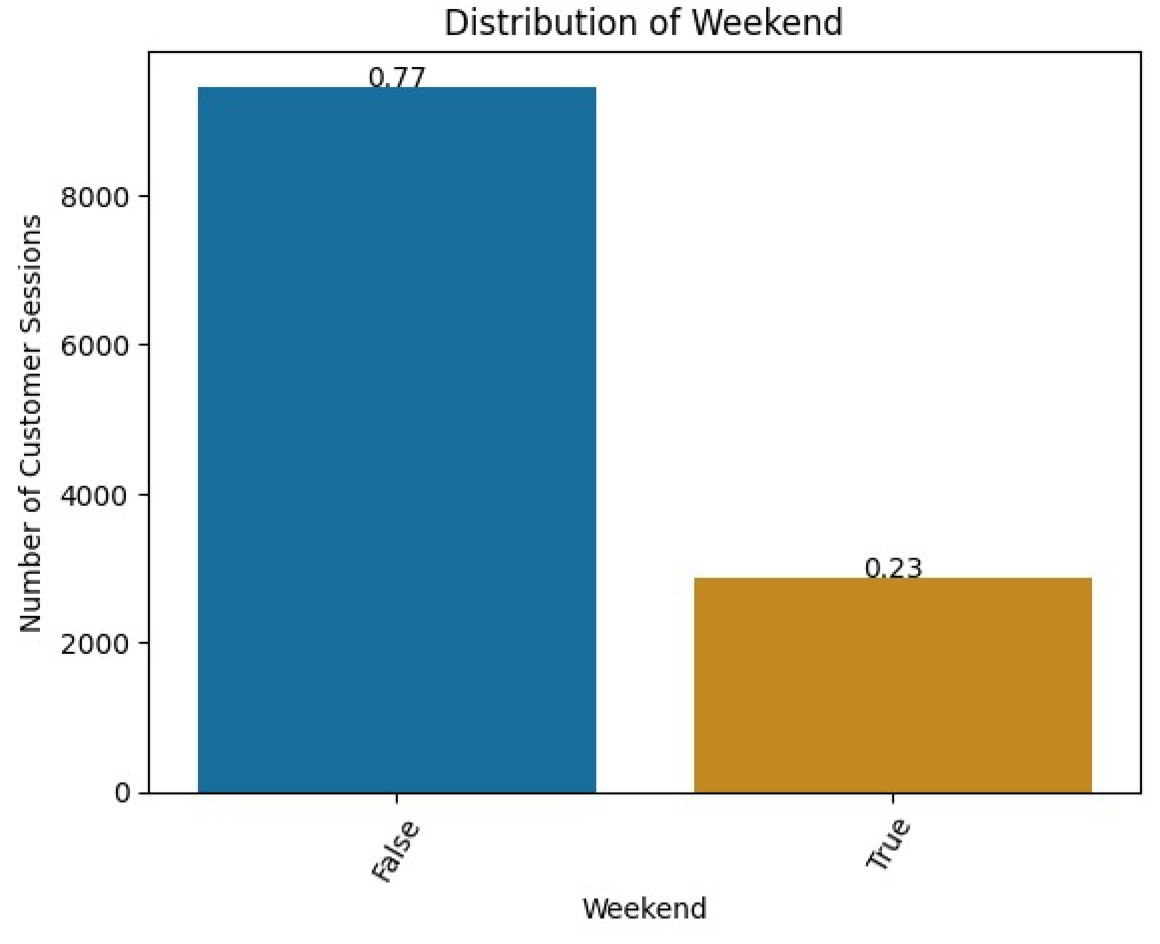
Visualizing categorical features (Traffic Type)



Visualizing categorical features (Visitor type)



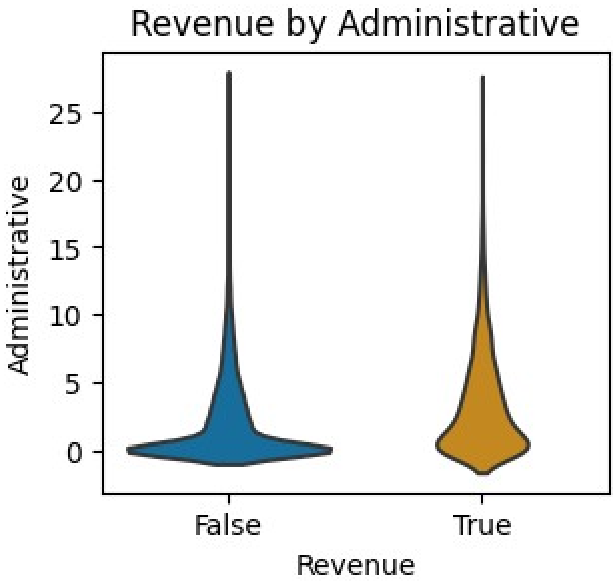
Visualizing categorical features (Weekend)

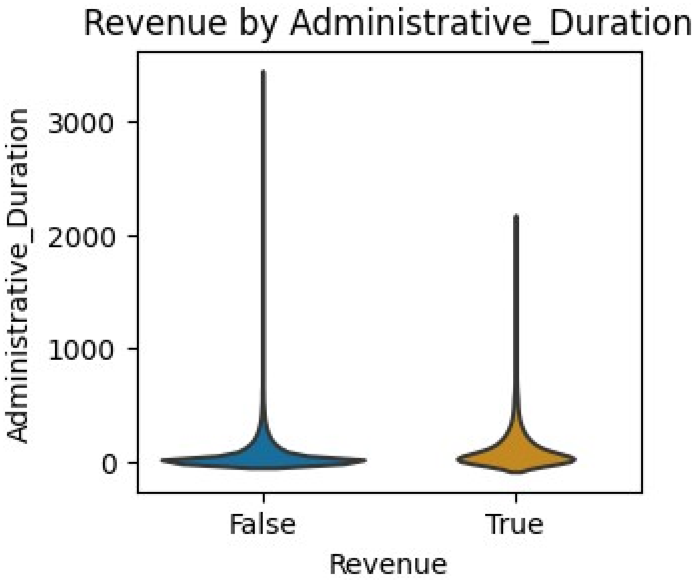


Knowledge Update

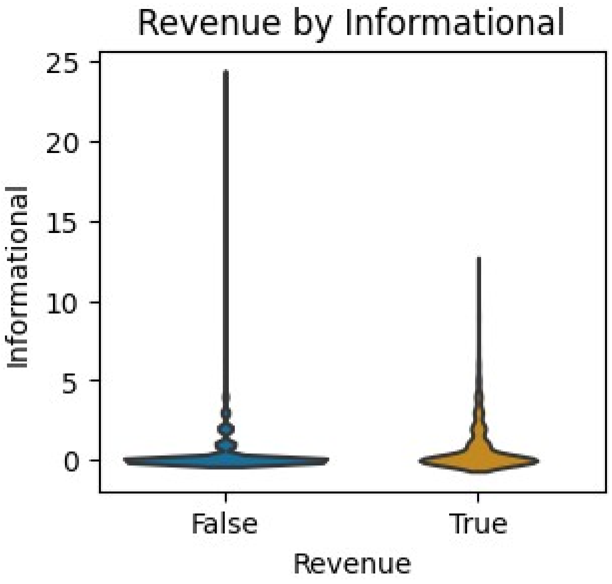
- There is clear imbalance in classes of independent variable
- Most customers shop in the months of May and November and do most of their shopping during the week. They use operating system 2, browser 2, and use traffic type 2. They live in region 1 and are returning customers. Most of the do not purchase anything.

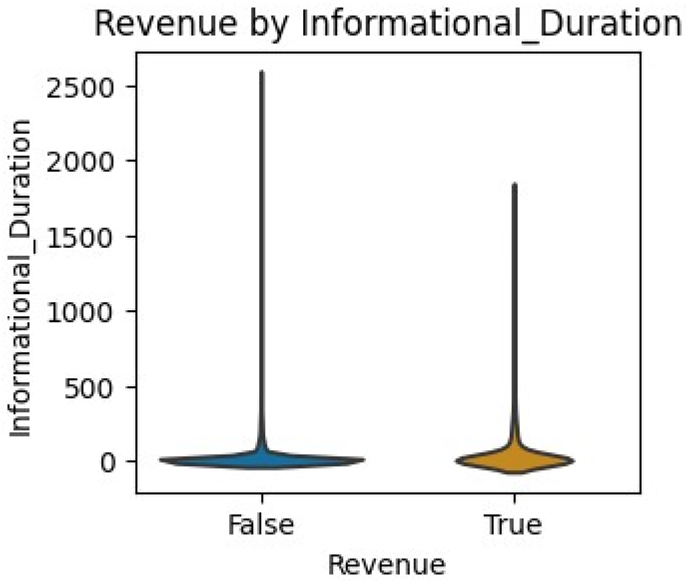
Visualizing numerical features (Administrative and its duration)



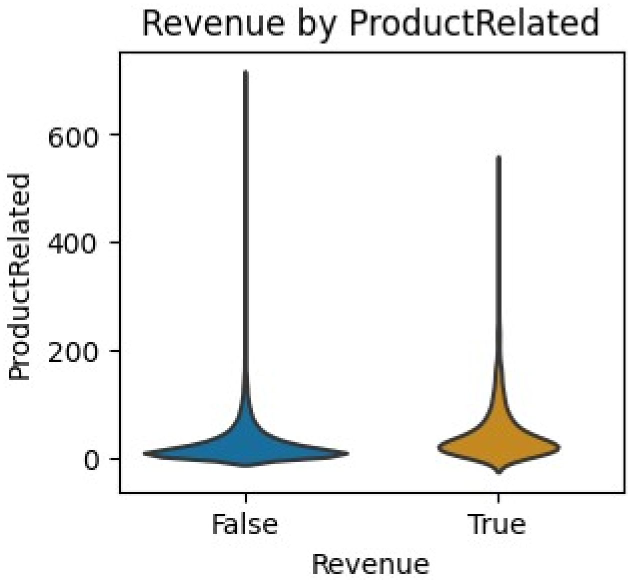


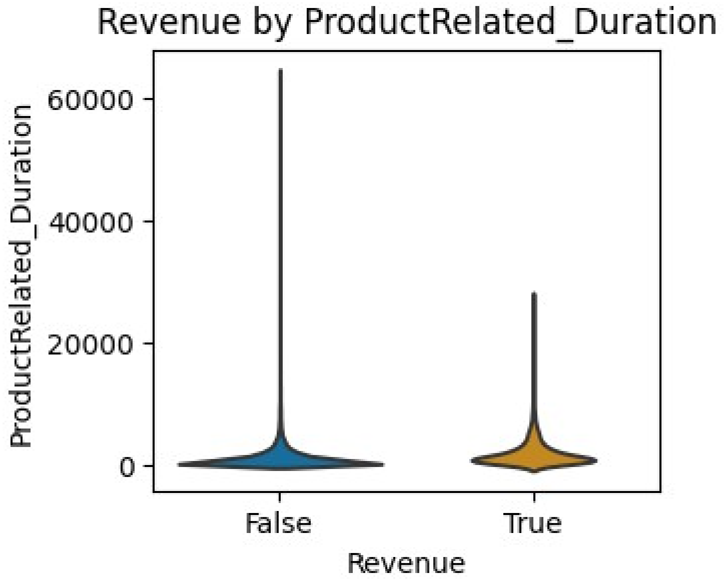
Visualizing numerical features (informational and its duration)



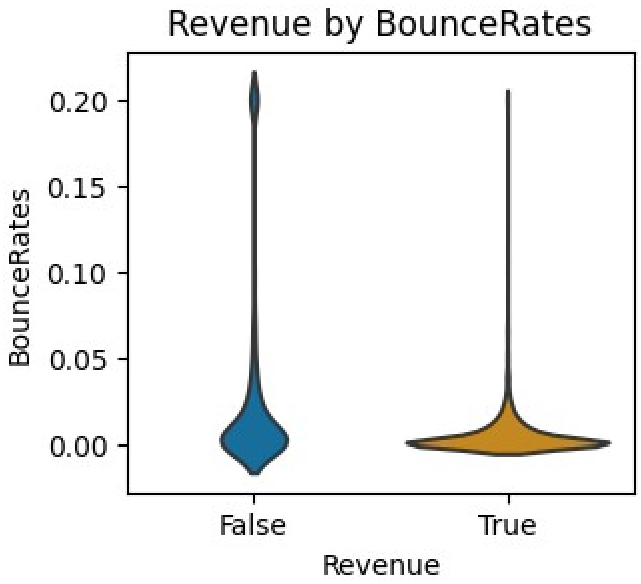


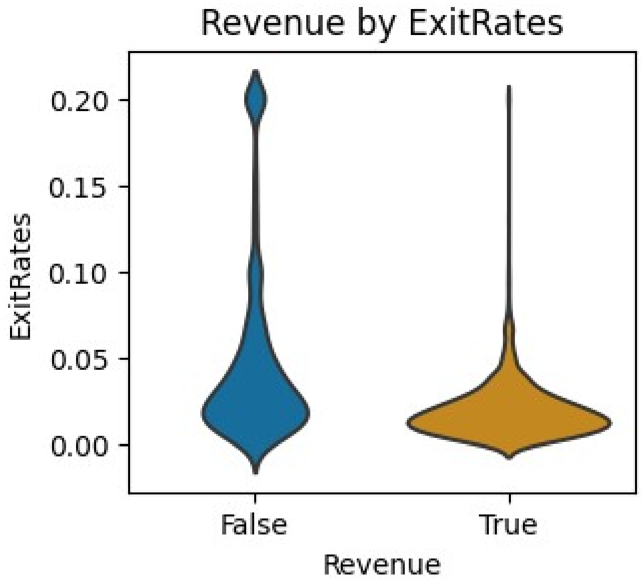
Visualizing numerical features (Product related and duration)



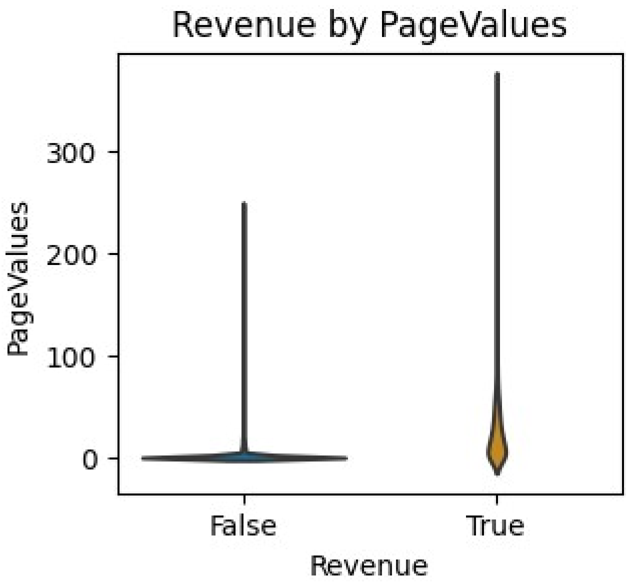


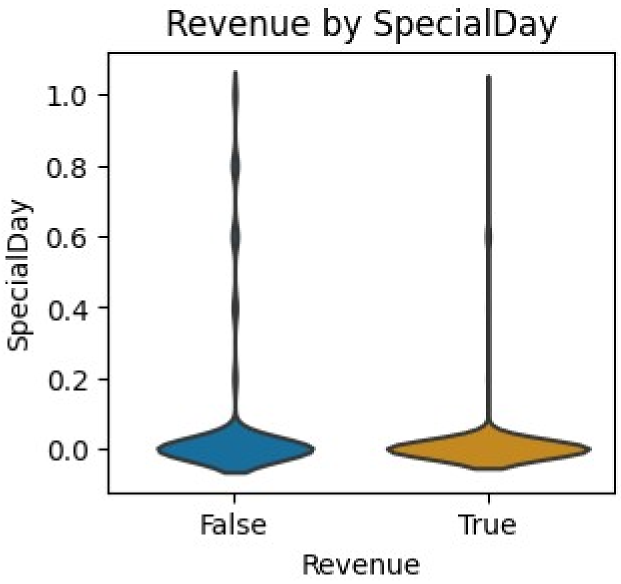
Visualizing numerical features (Bounce and exit rates)





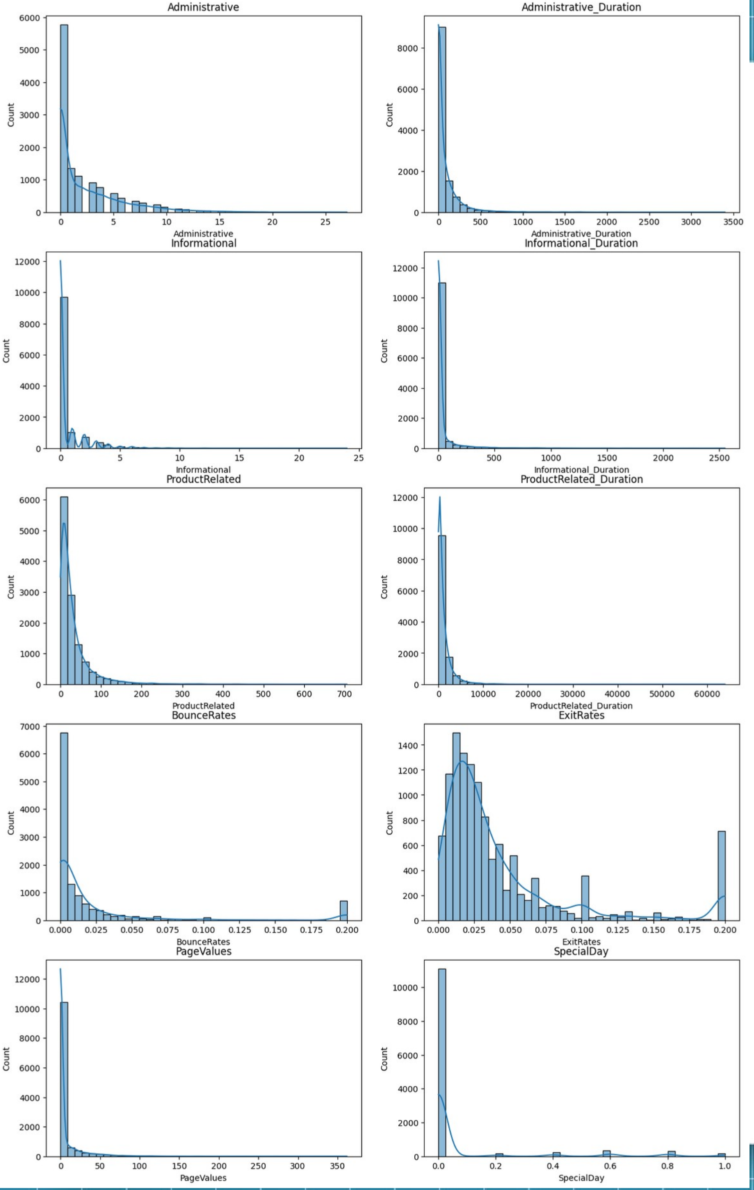
Visualizing numerical features (page values and special day)





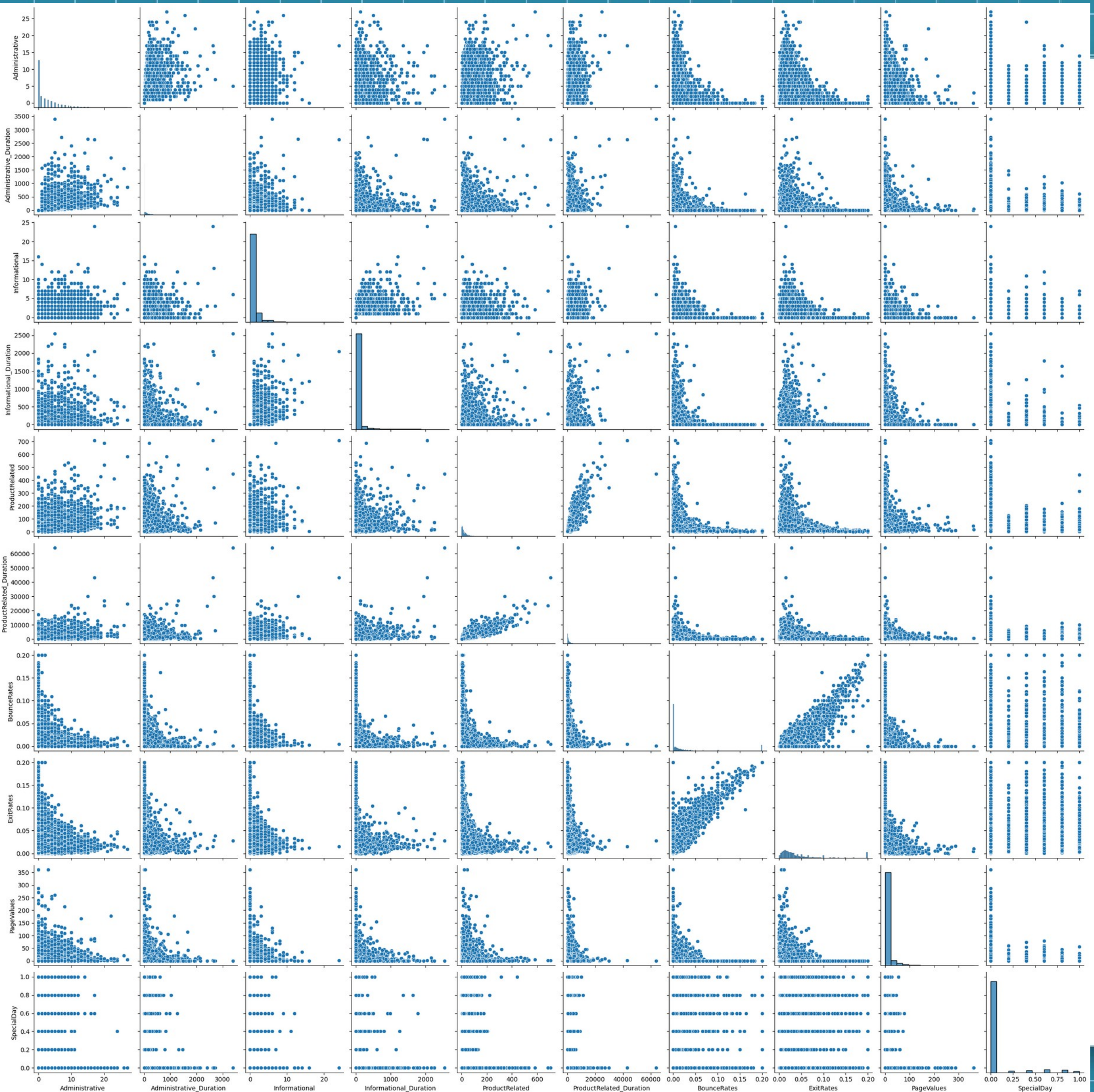
Knowledge Update

- Insights derived from the violin plots is: the higher the number of pages visited the more customers will purchase something.
- Another insight is customers who purchased an item have shorter exit rates and bounce rates than those customers that did not purchase an item.
- One strategy to keep customers on the website longer is providing items they are interested in purchasing through a recommender system.



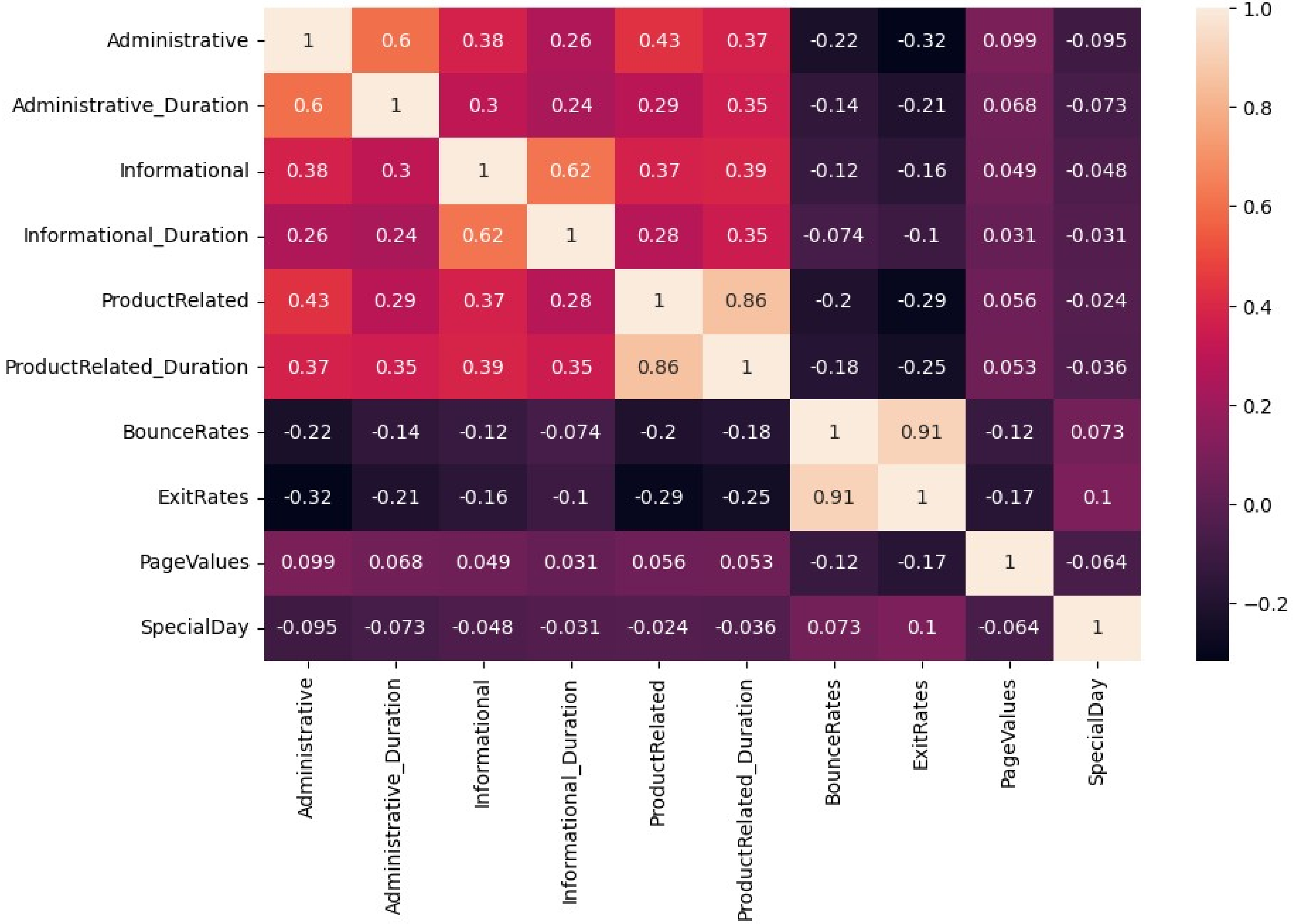
Distribution of Numeric Variables

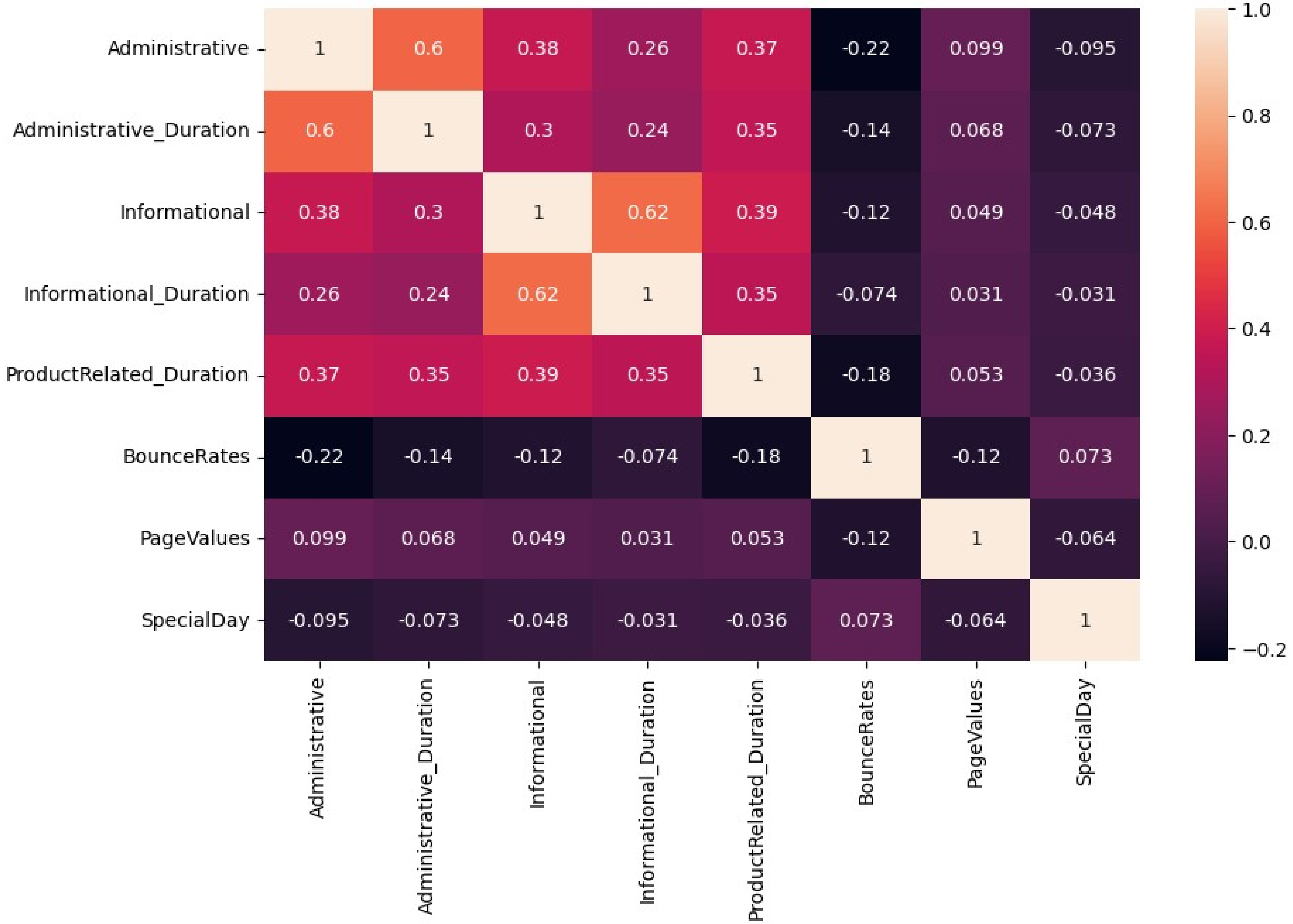
- Many distributions are skewed right
- Possible outliers

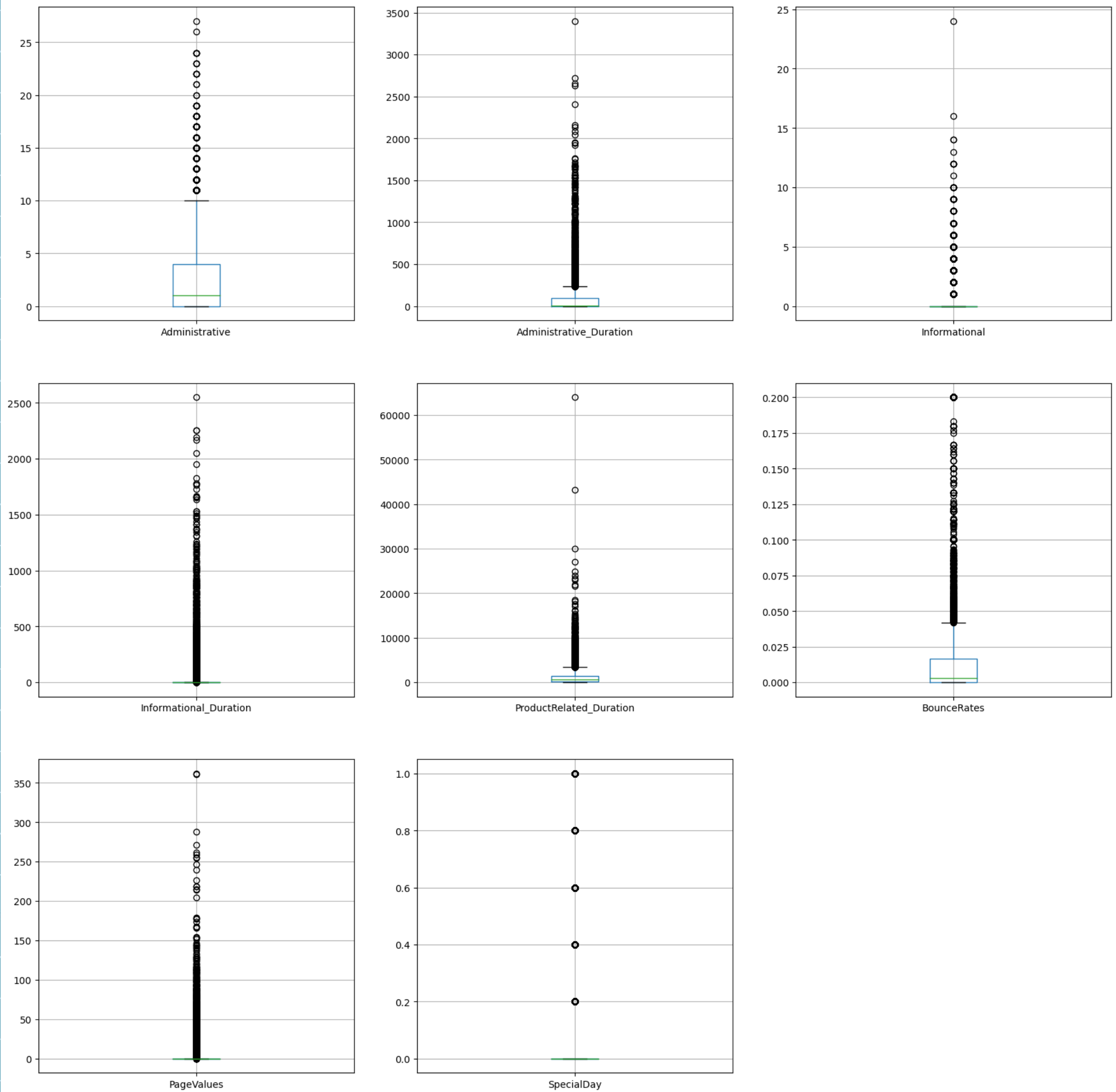


Pair plots

- Based on scatter plots there may be multicollinearity


Administrative
Administrative_Duration
Informational
Informational_Duration
ProductRelated
ProductRelated_Duration
BounceRates
ExitRates
PageValues
SpecialDay
1 0.6 0.38 0.26 0.43 0.37 -0.22 -0.32 0.099 -0.095
0.6 1 0.3 0.24 0.29 0.35 -0.14 -0.21 0.068 -0.073
0.38 0.3 1 0.62 0.37 0.39 -0.12 -0.16 0.049 -0.048
0.26 0.24 0.62 1 0.28 0.35 -0.074 -0.1 0.031 -0.031
0.43 0.29 0.37 0.28 1 0.86 -0.2 -0.29 0.056 -0.024
0.37 0.35 0.39 0.35 0.86 1 -0.18 -0.25 0.053 -0.036
-0.22 -0.14 -0.12 -0.074 -0.2 -0.18 1 0.91 -0.12 0.073
-0.32 -0.21 -0.16 -0.1 -0.29 -0.25 0.91 1 -0.17 0.1
0.099 0.068 0.049 0.031 0.056 0.053 -0.12 -0.17 1 -0.064
-0.095 -0.073 -0.048 -0.031 -0.024 -0.036 0.073 0.1 -0.064 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Administrative
Administrative_Duration
Informational
Informational_Duration
ProductRelated_Duration
BounceRates
PageValues
SpecialDay
1 0.6 0.38 0.26 0.37 -0.22 0.099 -0.095
0.6 1 0.3 0.24 0.35 -0.14 0.068 -0.073
0.38 0.3 1 0.62 0.39 -0.12 0.049 -0.048
0.26 0.24 0.62 1 0.35 -0.074 0.031 -0.031
0.37 0.35 0.39 0.35 1 -0.18 0.053 -0.036
-0.22 -0.14 -0.12 -0.074 -0.18 1 -0.12 0.073
0.099 0.068 0.049 0.031 0.053 -0.12 1 -0.064
-0.095 -0.073 -0.048 -0.031 -0.036 0.073 -0.064 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2



Outliers

- Administrative : 404 and 3.27%
- Administrative_Duration : 1172 and 9.50%
- Informational : 2631 and 21.33%
- Informational_Duration : 2405 and 19.50%
- ProductRelated_Duration : 961 and 7.79%
- BounceRates : 1551 and 12.57%
- PageValues : 2730 and 22.14%
- SpecialDay : 1251 and 10.14%

Outliers detection and resolution

- q75, q25 = np.percentile(df['Administrative'], [75, 25])
- iqr = q75 – q25
- min_val = q25 – (iqr*1.5)
- max_val = q75 + (iqr*1.5)

- StandardScaler for numeric data
- OneHotEncoder for categorical data

Logistic Regression

Model score: 0.880

Confusion Matrix:

```
[[2033   51]
 [ 246  136]]
```

Classification Report:

	precision	recall	f1-score	support
False	0.89	0.98	0.93	2084
True	0.73	0.36	0.48	382
accuracy			0.88	2466
macro avg	0.81	0.67	0.70	2466
weighted avg	0.87	0.88	0.86	2466

SVC(C=0.025, probability=True)

Model score: 0.870

Confusion Matrix:

[[2051   33]

 [288   94]]

Classification Report:

	precision	recall	f1-score	support
False	0.88	0.98	0.93	2084
True	0.74	0.25	0.37	382
accuracy			0.87	2466
macro avg	0.81	0.62	0.65	2466
weighted avg	0.86	0.87	0.84	2466

DecisionTreeClassifier

Model score: 0.859

Confusion Matrix:

```
[[1912  172]
 [ 176  206]]
```

Classification Report:

	precision	recall	f1-score	support
False	0.92	0.92	0.92	2084
True	0.54	0.54	0.54	382
accuracy			0.86	2466
macro avg	0.73	0.73	0.73	2466
weighted avg	0.86	0.86	0.86	2466

RandomForestClassifier

Model score: 0.895

Confusion Matrix:

[[2018   66]
 [192 190]]

Classification Report:

	precision	recall	f1-score	support
False	0.91	0.97	0.94	2084
True	0.74	0.50	0.60	382
accuracy			0.90	2466
macro avg	0.83	0.73	0.77	2466
weighted avg	0.89	0.90	0.89	2466

•

GradientBoostingClassifier

Model score: 0.903

Confusion Matrix:

[[2005   79]
 [160 222]]

Classification Report:

	precision	recall	f1-score	support
False	0.93	0.96	0.94	2084
True	0.74	0.58	0.65	382
accuracy			0.90	2466
macro avg	0.83	0.77	0.80	2466
weighted avg	0.90	0.90	0.90	2466

MLPClassifier

Model score: 0.873

Confusion Matrix:

[[1941 143]
 [170 212]]

Classification Report:

	precision	recall	f1-score	support
False	0.92	0.93	0.93	2084
True	0.60	0.55	0.58	382
accuracy			0.87	2466
macro avg	0.76	0.74	0.75	2466
weighted avg	0.87	0.87	0.87	2466

Limitations

- Ignores factors like product features/price
- No data on past purchase
- No data on product review

Conclusion

- Through this process, we noticed Gradient Boosting Classifier algorithm performed better than all the others with 0.90 model score(f1 score)
- Gradient Boosting Classifier algorithm can be preferred in class imbalance problem

References

- Peighambari, K., Sattari, S., Kordestani, A., & Oghazi, P. (2016). Consumer Behavior Research. SAGE Open, 6(2), 215824401664563. https://doi.org/10.1177/2158244016645638
- Rojhe, K. C. (2020). Review Paper on Factors Influencing Consumer Behavior. ResearchGate. https://www.researchgate.net/publication/342876391_Review_Paper_on_Factors_Influencing_Consumer_Behavior
- Chovanová, H. H., Korshunov, A., & Babčanová, D. (2015). Impact of Brand on Consumer Behavior. Procedia. Economics and Finance, 34, 615–621. https://doi.org/10.1016/s2212-5671(15)01676-7
- Sv, S. (2022). A STUDY ON CONSUMER BEHAVIOUR TOWARDS ONLINE SHOPPING. ResearchGate. https://www.researchgate.net/publication/358605912_A_STUDY_ON_CONSUMER_BEHAVIOUR_TOWARDS_ONLINE_SHOPPING
- Alshweesh, R., & Bandi, S. (2022). The Impact of E-Commerce on Consumer Purchasing Behavior: The Mediating Role of Financial Technology. International Journal of Research and Review, 9(2), 479–499. https://doi.org/10.52403/ijrr.20220261

Thank You !

- Again, for code and more details: binabh.com.np/code.html