ILLINOIS INSTITUTE OF TECHNOLOGY

Analysis on Online Shoppers Purchasing Intention

Under the guidance of
Professor - Jawahar Panchal

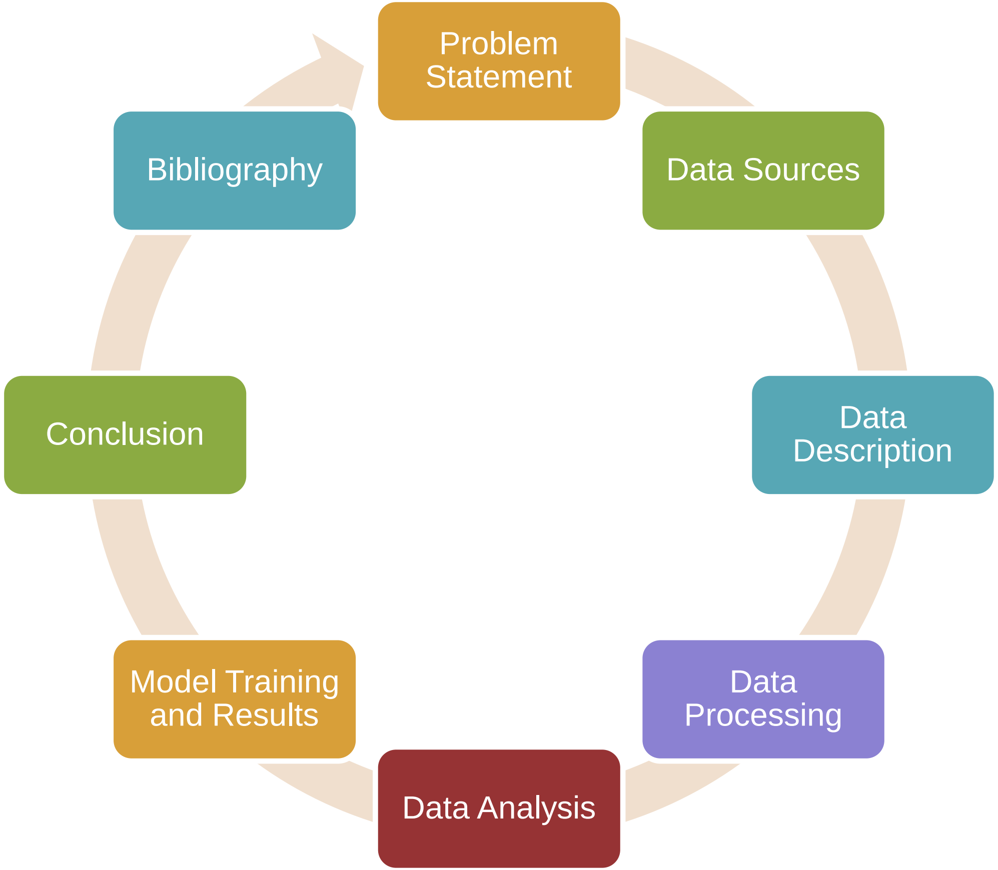
CSP571 Data Preparation and Analysis

Team Structure

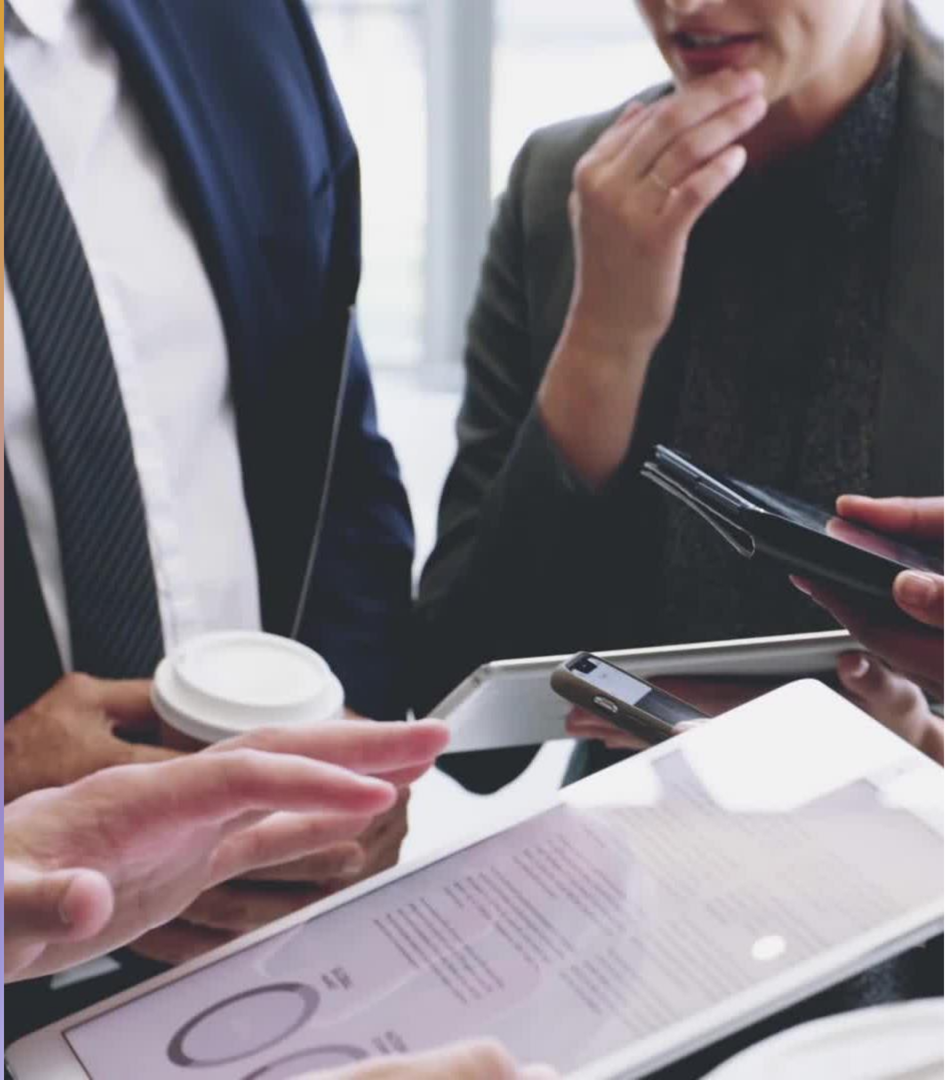
- Divya Sai Sree Chintala
- Team Leader: A20561001

- Sneha Joshi
- Team member: A20540613

Project Outline



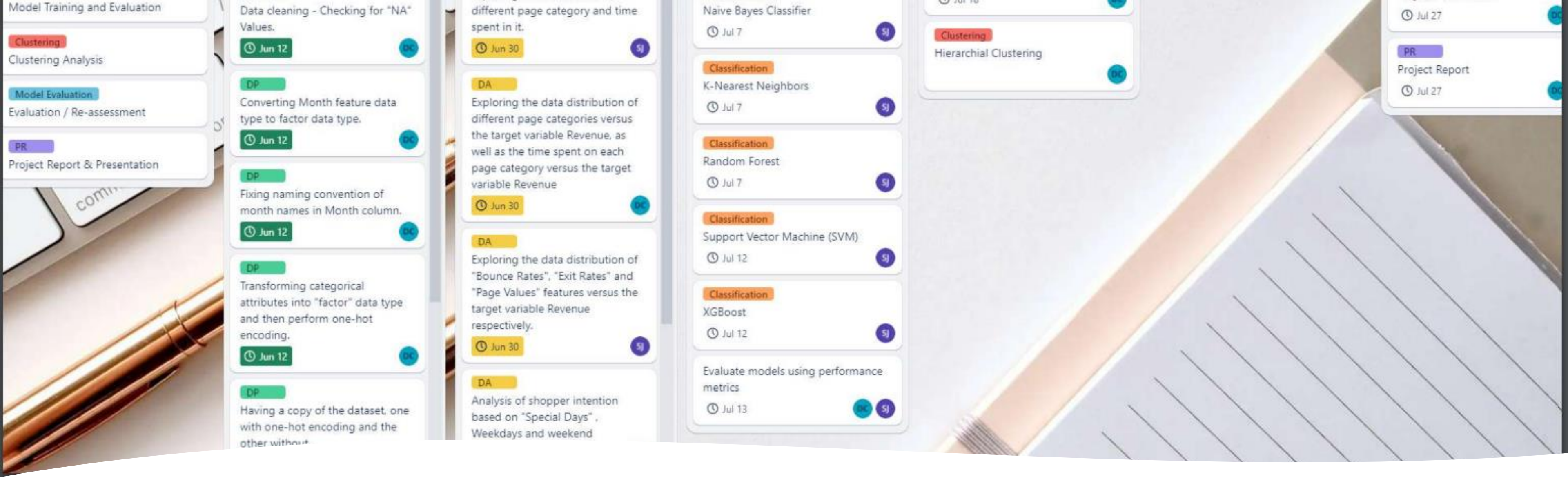
Problem Statement



Problem Statement

- Analyze trends in the online shoppers purchasing intention dataset using exploratory data analysis techniques and build machine learning models to predict the purchasing intentions of visitors to a store's website both using supervised and un-supervised techniques.

Project Planning



Project Planning and execution

- Used Trello for the planning and keeping track of the project.
- Logistic Regression and GMM were not implemented and put to the backlog, will be included in the future work.

ILLINOIS INSTITUTE OF TECHNOLOGY

Workflow overview

Data Collection → Data Preparation & Analysis → Modelling → Results & Conclusion

Data Collection

- Online Shopper Purchase Intention Dataset from UCI

Data Preparation & Analysis

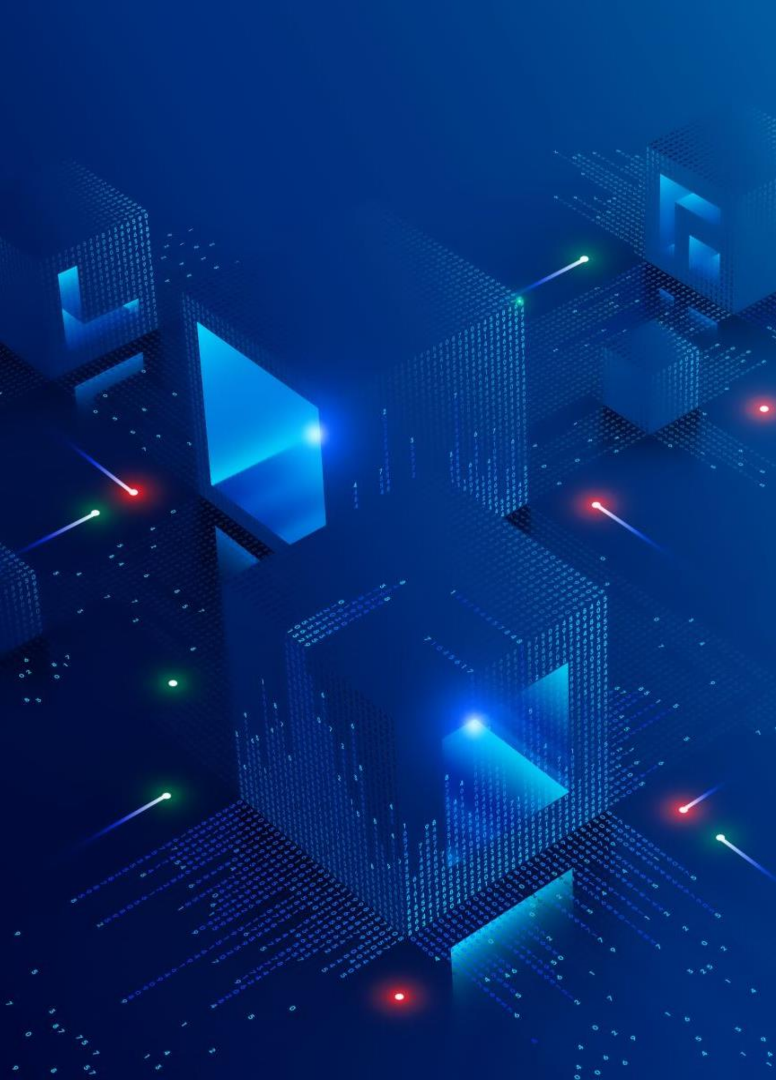
- Data cleaning.
- Data Balancing.
- Univariate, Bivariate and Multivariate analysis.

Modelling

- Supervised Modeling
  - Naive Bayes
  - K-NN
  - Random Forest
  - SVM
  - XGBoost
- Un-supervised
  - K-means Clustering
  - DBScan Clustering
  - Hierarchical clustering

Results & Conclusion

- Analysed conclusion.
- Model reports and evaluation.
- Future work.



Data Sources

- The data that is being used in this project was obtained from the UC Irvine Machine Learning Repository.

- **Data set contributors:**

  - **C. Okan Sakar :** Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Bahcesehir University, 34349 Besiktas, Istanbul, Turkey

  - **Yomi Kastro:** Inveon Information Technologies Consultancy and Trade, 34335 Istanbul, Turkey

Data description

- The dataset consists of feature vectors belonging to 12,330 sessions.
- The dataset consists of both numerical and categorical attributes. The 'Revenue' attribute can be used as the class label.

Attributes	
Administrative	Administrative Duration
Informational	Informational Duration
Product Related	Product Related Duration
Bounce rate	Exit rate
Page value	Special day
Operating system	Browser
Region	Traffic type
Visitor type	Weekend
Month of the year	Revenue

Data Preprocessing

Data processing

- Check number of observations with NA values
- Fixing naming convention of month names in Month column “June” ->”Jun”
- Convert Month feature data type to factor data type
- Transforming categorical attributes(OperatingSystems, Browser, Region, TrafficType, VisitorType) into “factor” data type and then perform one-hot encoding
- Convert Revenue attribute data type to a factor.
- Transforming Boolean attributes(Weekend, Revenue) into “int” data type
- Train - Test split : 70:30 split
- One hot encoding of train and test set

Data Balancing

There is huge imbalance in data set as Revenue=0 is the majority. Hence the algorithm tries to over fit on majority class.

Number of observations with Revenue as False = 10422

Number of observations with Revenue as True = 1908

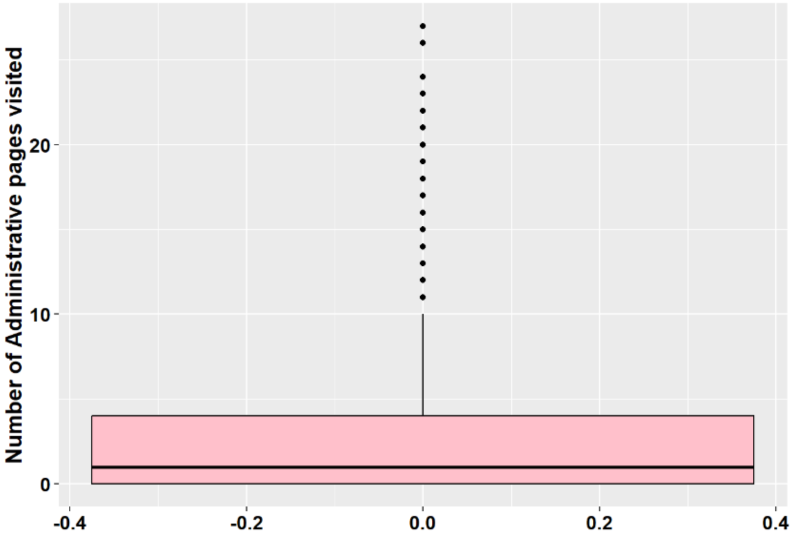
Here we are trying to increase minority class observations using SMOTE(Synthetic Minority Over-sampling Technique) algorithm.

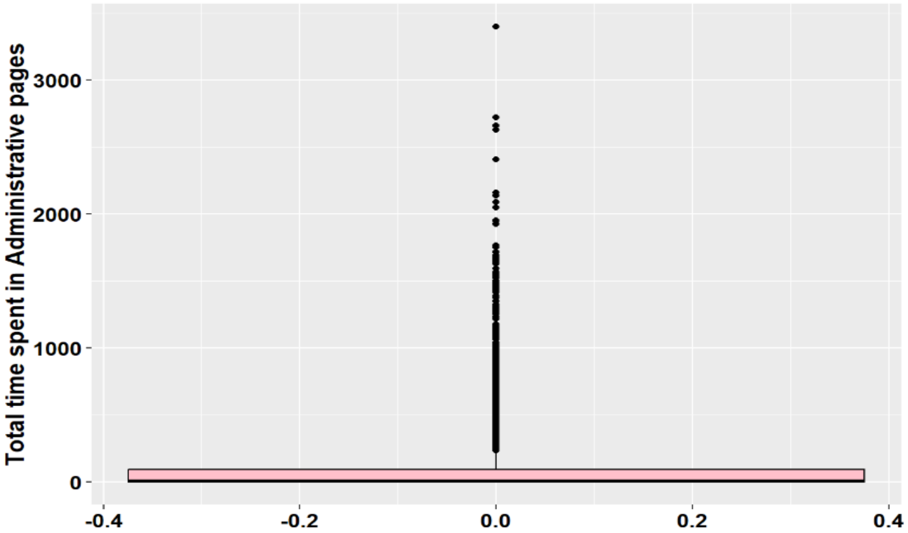
Data Analysis

Data Analysis

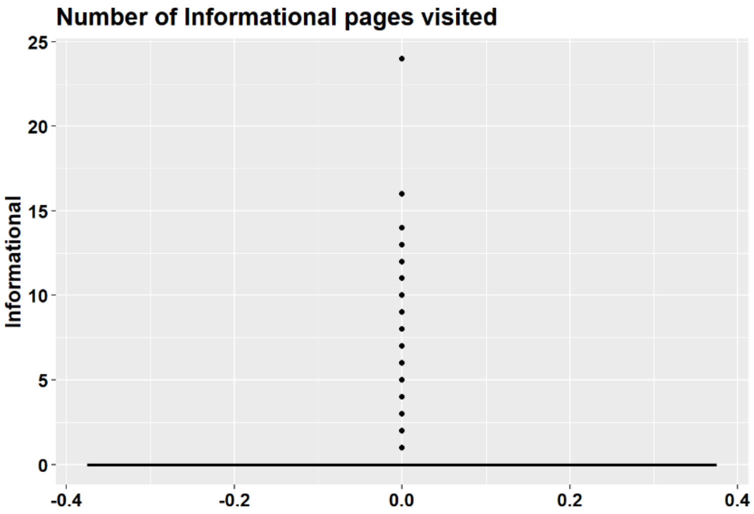
1) Exploring data distribution of different page category and time spent in it.

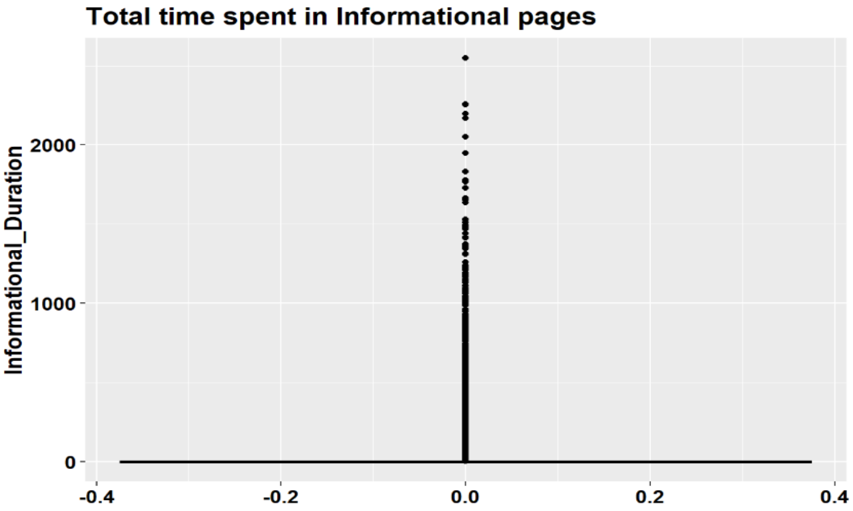
a) Exploring data pattern of “Administrative” and “Administrative_Duration”



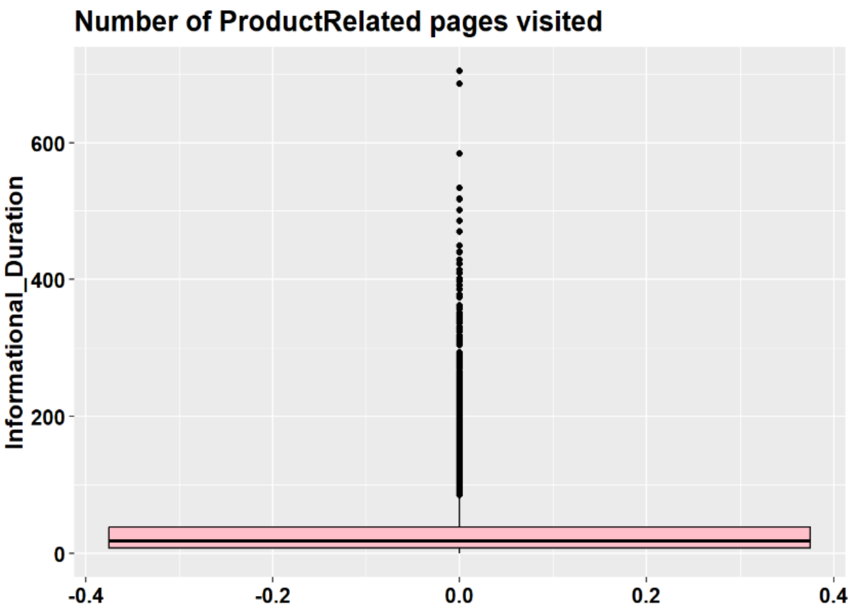


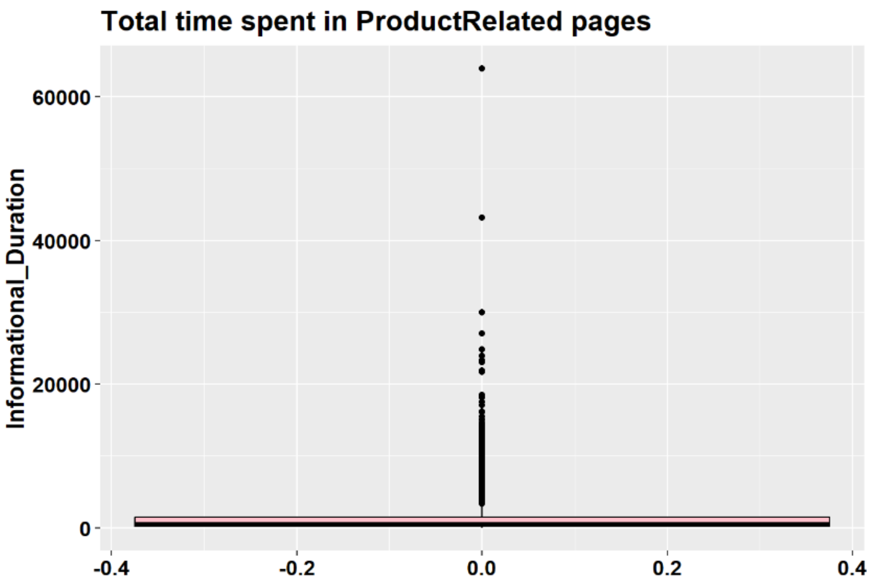
b) Exploring data pattern of “Informational” and “Informational_Duration”





c) Exploring data pattern of “Product Related” and “Product Related Duration”



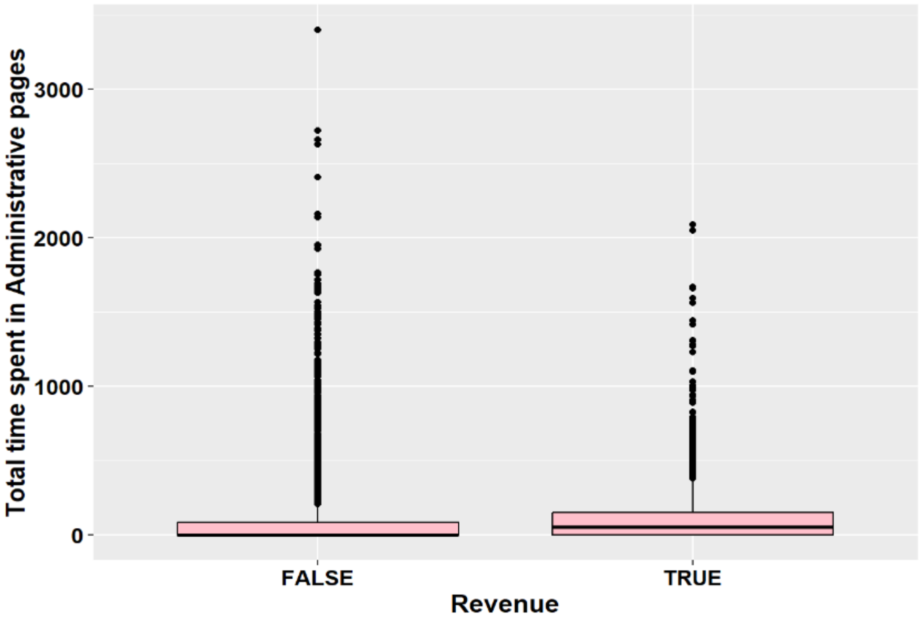


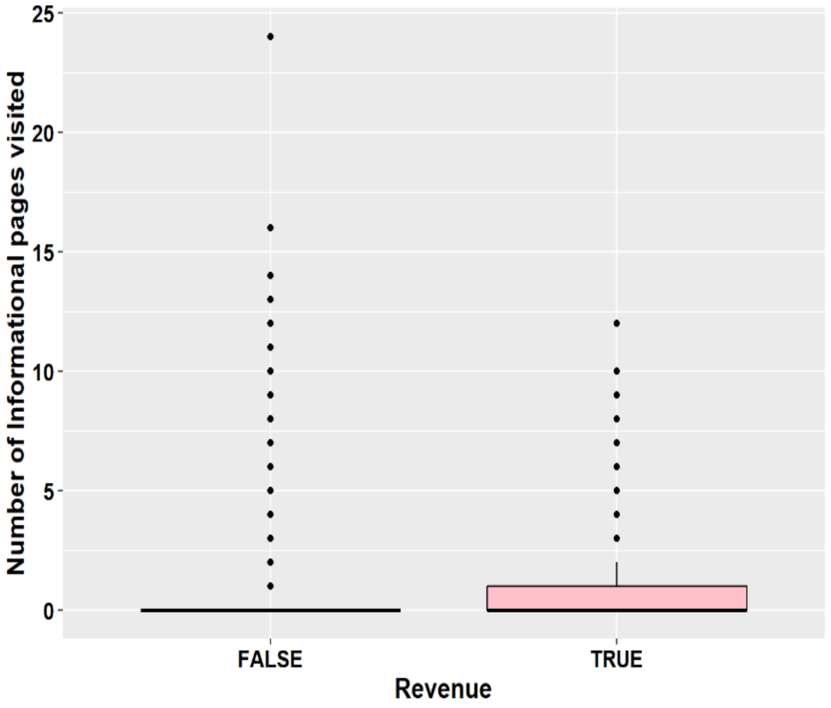
Summary

Analyzing number of page visit of 3 different page categories it clearly says that customers are interested more in Product related pages rather than knowing information of the product in detail.

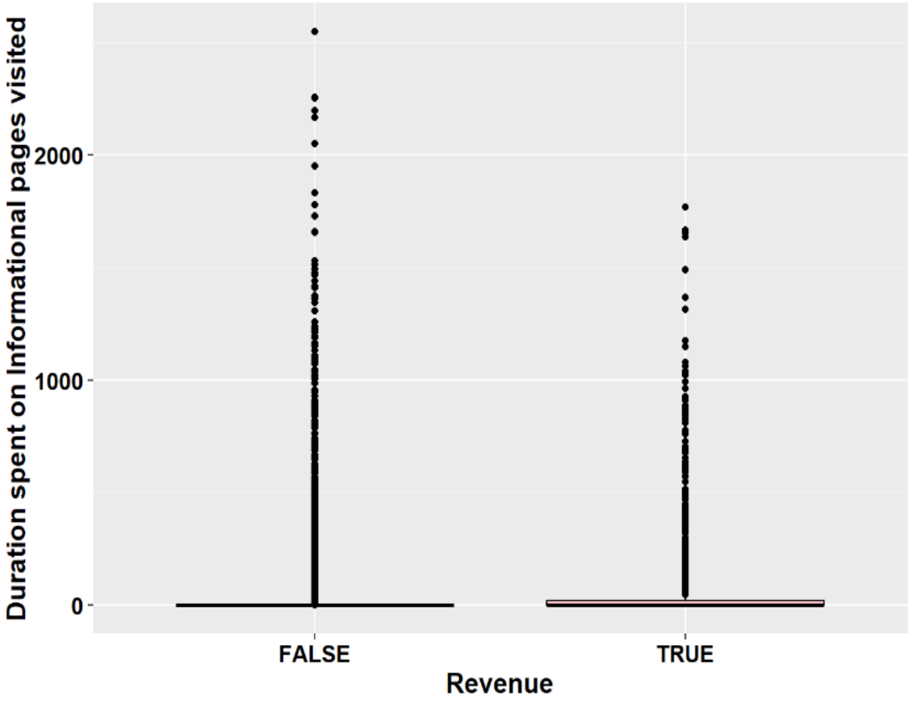
Analyzing total time spent in 3 different page categories, it clearly says that customers spend most of the time in product related pages whereas they are not interested in spending time in information related pages.

2) Exploring the data distribution of different page categories versus the target variable Revenue, as well as the time spent on each page category versus the target variable Revenue.

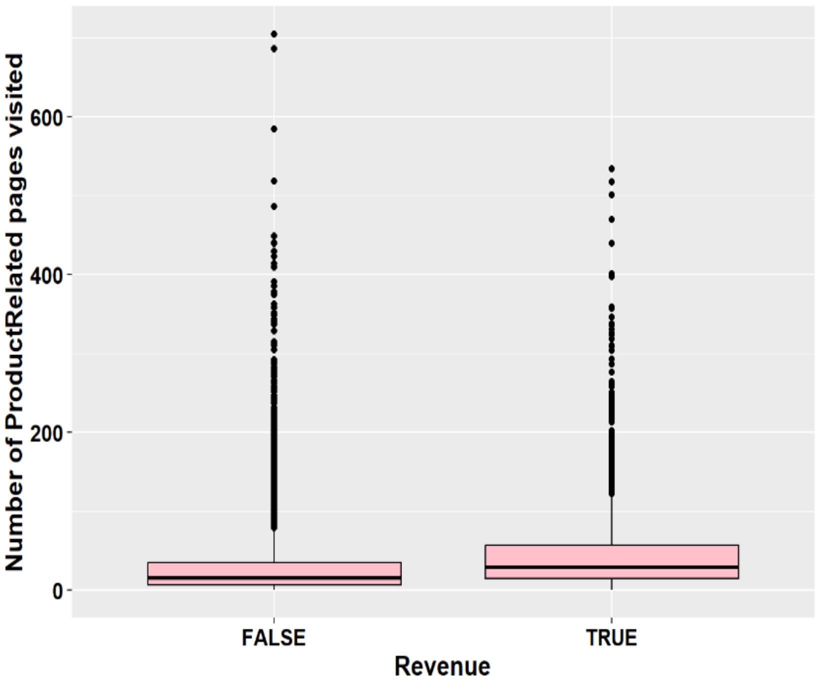




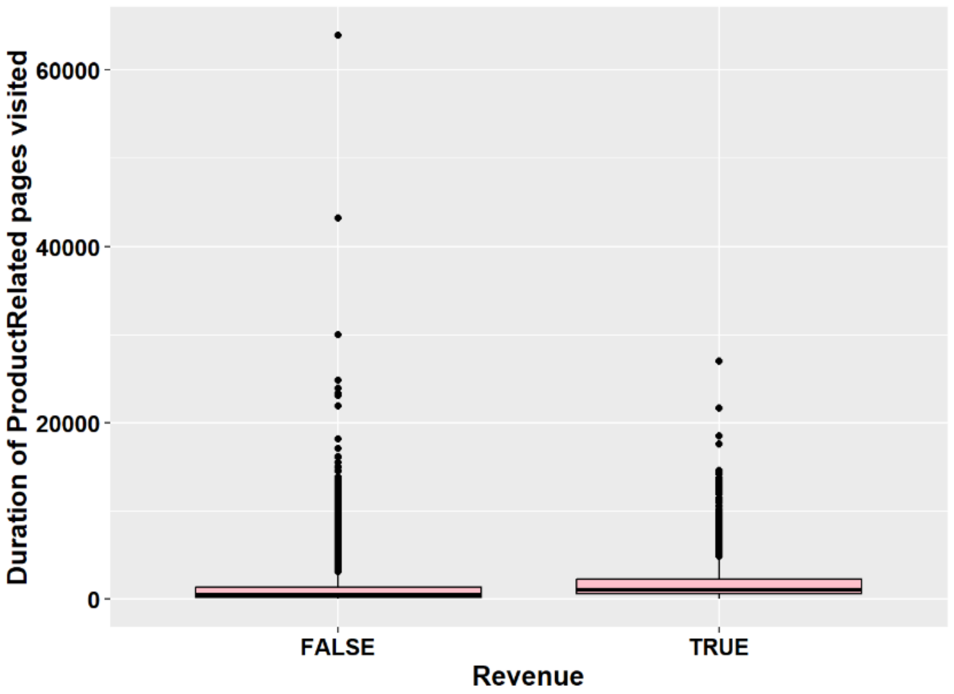
Number of Informational pages visited
25
20
15
10
5
0
FALSE
TRUE
Revenue



Duration spent on Informational pages visited
2000
1000
0
FALSE
TRUE
Revenue



Number of ProductRelated pages visited
600
400
200
0
FALSE
TRUE
Revenue



Duration of ProductRelated pages visited
60000
40000
20000
0
FALSE
TRUE
Revenue

Summary

People who end up buying will mostly visit administrative page and spend almost 52seconds.

People who end up not buying will mostly not visit administrative page.

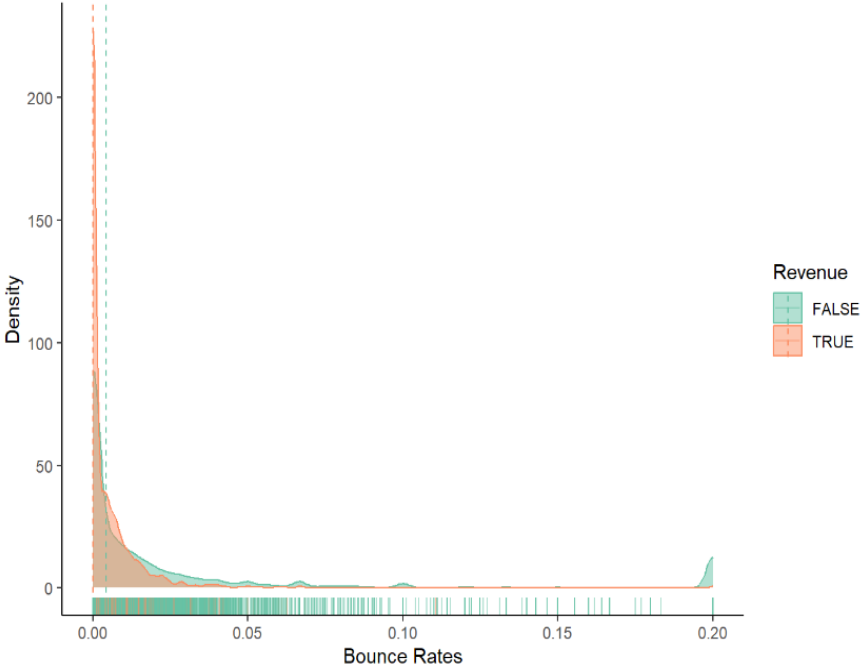
People are least interested in visiting informational page.

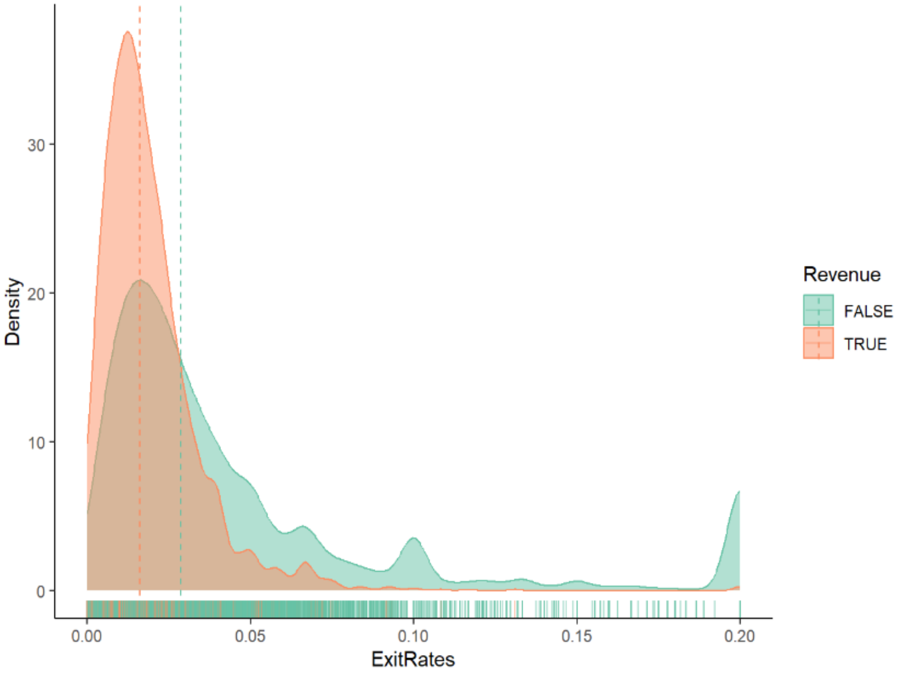
People who end up buying will mostly visit product related page and spend almost 1109 seconds.

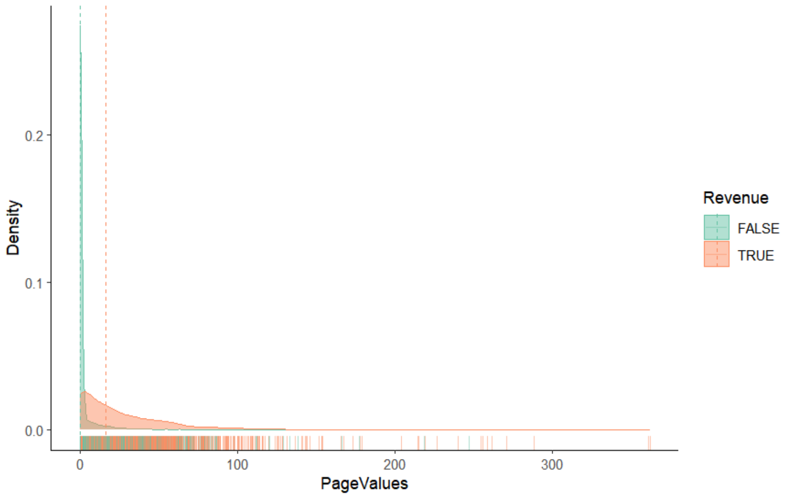
People who will end up buying will mostly visit product related page and spend almost 510 seconds.

But people who end up buying will visit more product related than the ones who don't.

3) “<u>Bounce Rates”, “Exit Rates” and “Page Values”</u> features versus the target variable Revenue, respectively.



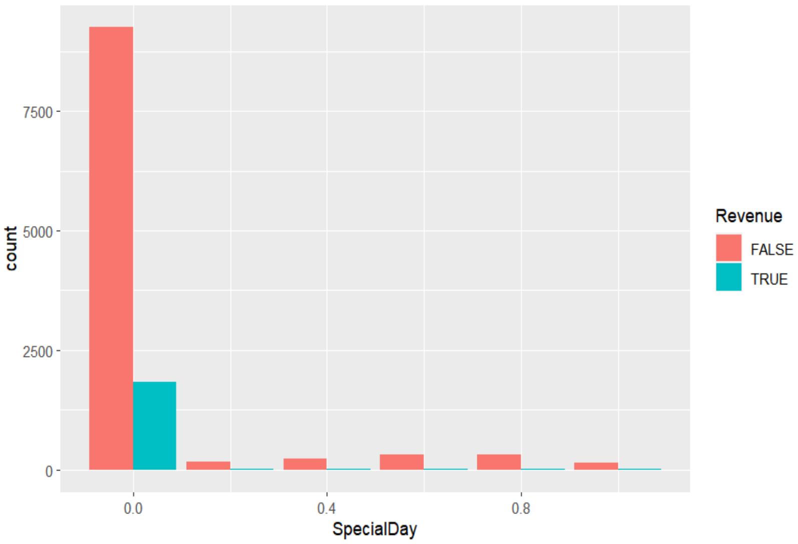




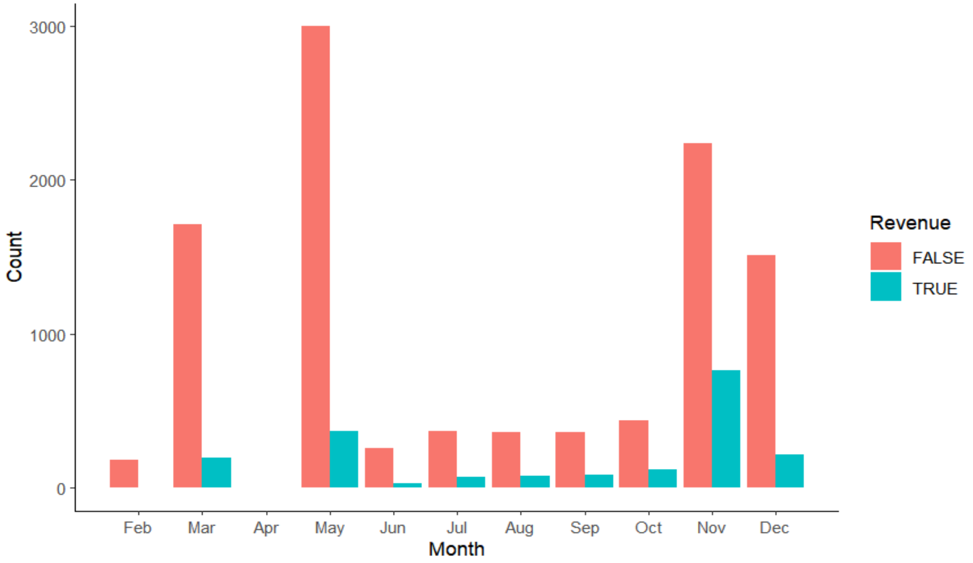
Summary

- There is no noticeable disparity in Bounce Rates between customers who made a purchase and those who did not.
- However, customers who ended up making a purchase had lower Exit Rates on average, indicating that they were more likely to remain on the website's pages.
- Additionally, customers who did not make a purchase had significantly lower Page Values, suggesting that they spent less time on related pages.

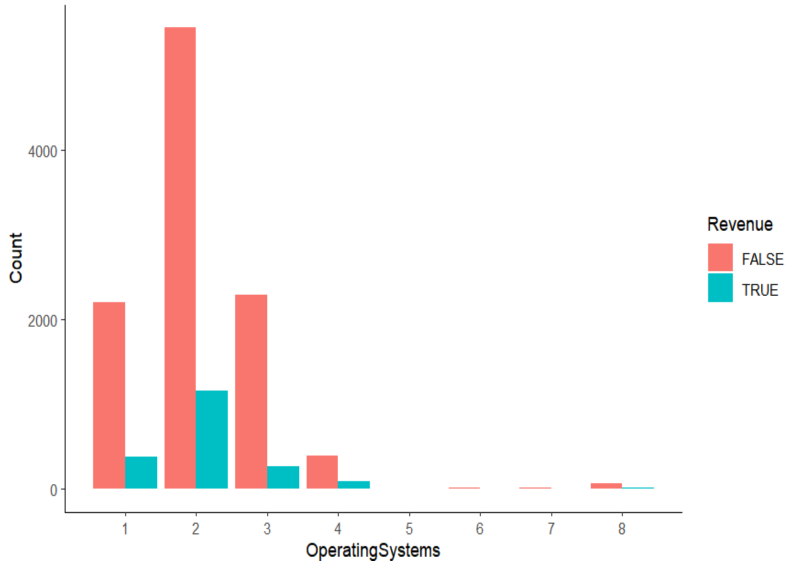
4) “Special Day” features versus the target variable Revenue



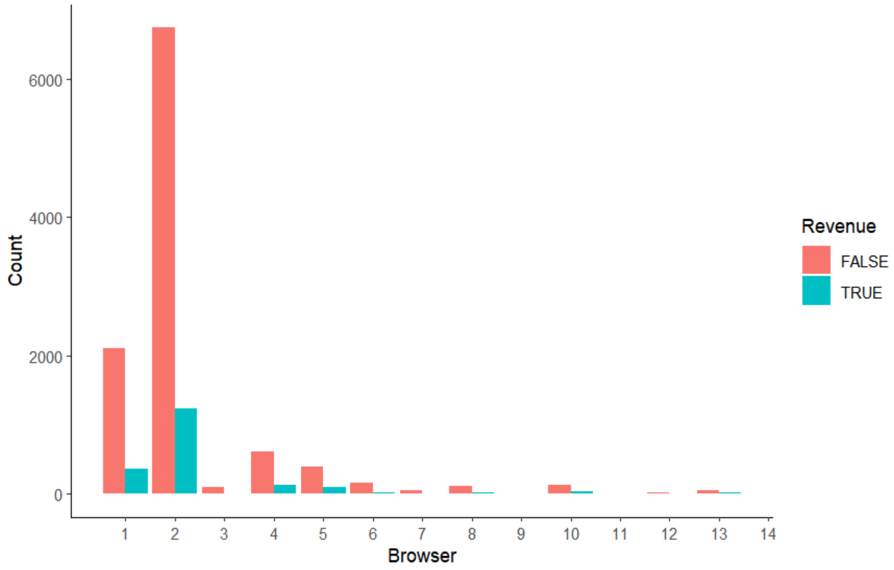
5) “Month” features versus the target variable Revenue.



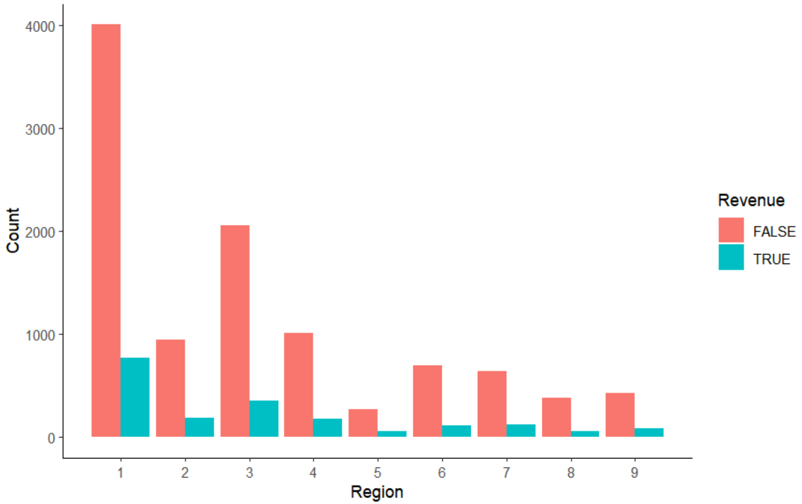
6) “Operating Systems” features versus the target variable Revenue.



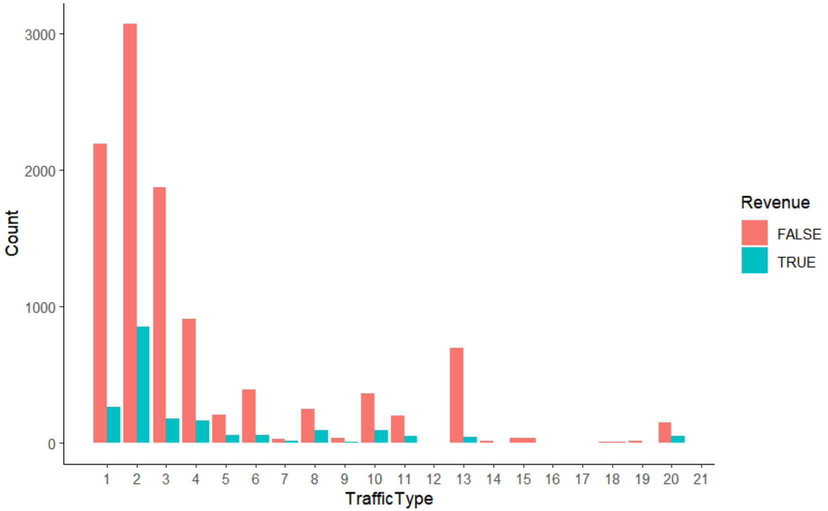
7) “Browser” features versus the target variable Revenue.



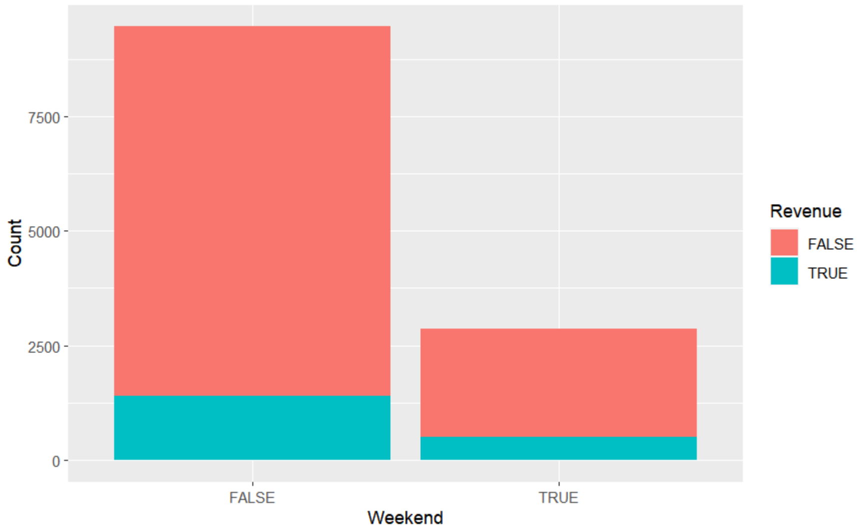
8) "Region" features versus the target variable Revenue.



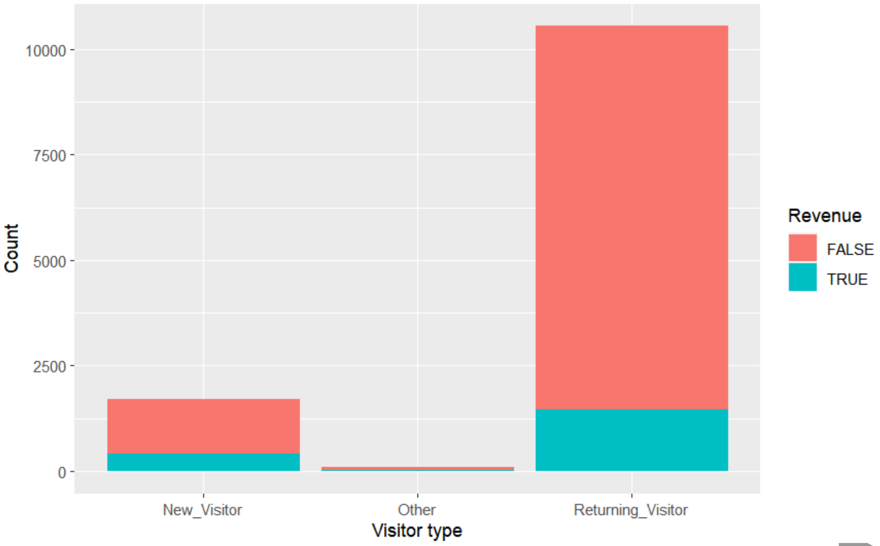
9) "Traffic Type" features versus the target variable Revenue.



10) “Weekend” features versus the target variable Revenue.



11) “Visitor Type” features versus the target variable Revenue.



Supervised Modeling

Model Training and Results

1) Naive Bayes

Results:

One hot encoding data

Average accuracy: 75.1%

Prediction	Reference	
	0	1
0	31.6	6.0
0	18.4	44.0

Data without one-hot encoding

Average accuracy: 84.5%

Prediction	Reference	
	0	1
0	84.5	15.5
0	0	0

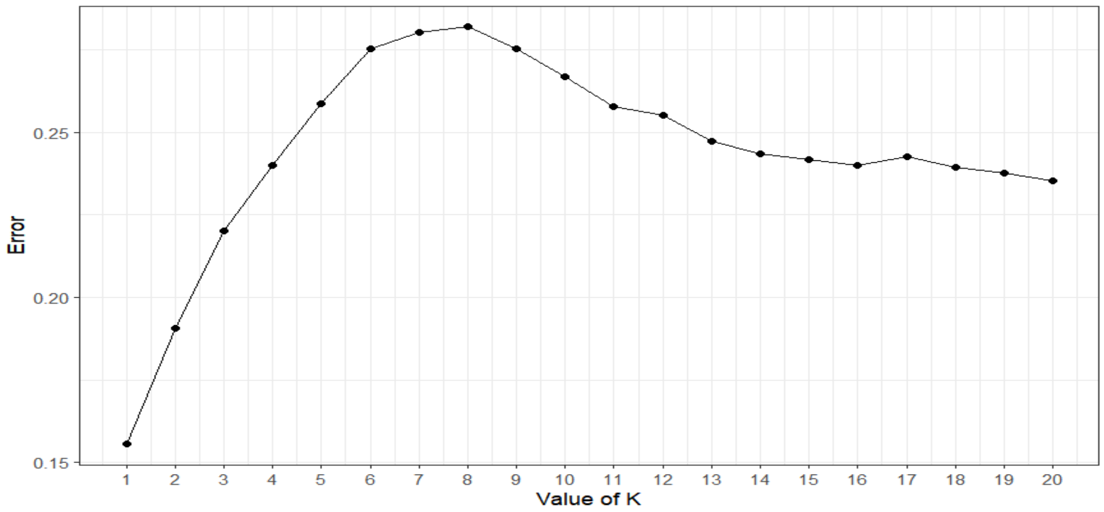
2) K-Nearest Neighbor

Trained on one-hot encoded dataset

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2847  296
         1  280  277

               Accuracy : 0.8443
                 95% CI : (0.8322, 0.8559)
    No Information Rate : 0.8451
    P-Value [Acc > NIR] : 0.5652

                  Kappa : 0.3984

 Mcnemar's Test P-Value : 0.5320

            Sensitivity : 0.9105
            Specificity : 0.4834
         Pos Pred Value : 0.9058
         Neg Pred Value : 0.4973
             Prevalence : 0.8451
         Detection Rate : 0.7695
   Detection Prevalence : 0.8495
      Balanced Accuracy : 0.6969

       'Positive' Class : 0
```

Visualizing KNN with different K values(number of nearest neighbor)

- Lowest error with k=1



3) Random Forest

Results:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2955  189
         1  172  384

               Accuracy : 0.9024
                 95% CI : (0.8924, 0.9118)
    No Information Rate : 0.8451
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6227

 Mcnemar's Test P-Value : 0.3997

            Sensitivity : 0.9450
            Specificity : 0.6702
         Pos Pred Value : 0.9399
         Neg Pred Value : 0.6906
             Prevalence : 0.8451
         Detection Rate : 0.7986
   Detection Prevalence : 0.8497
      Balanced Accuracy : 0.8076

       'Positive' Class : 0

Thus accuracy of random forest is 89.81%
```

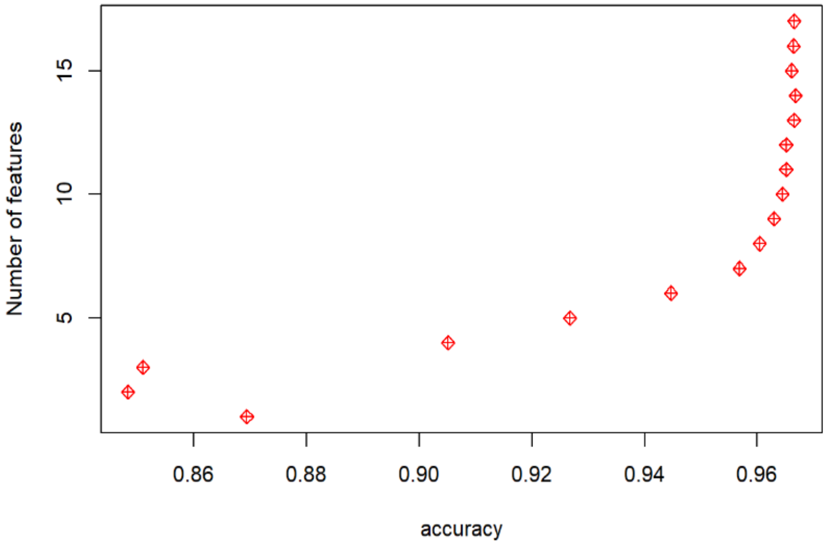
Recursive Feature Elimination

	Attribute	Accuracy
14	Region	0.9667576
17	Weekend	0.9665521
13	Browser	0.9664835
16	VisitorType	0.9664151
15	TrafficType	0.9660722
12	OperatingSystems	0.9651126
11	Month	0.9651125
10	SpecialDay	0.9644273
9	PageValues	0.9629195
8	ExitRates	0.9603834
7	BounceRates	0.9568190
6	ProductRelated_Duration	0.9446862
5	ProductRelated	0.9267302
4	Informational_Duration	0.9051401
1	Administrative	0.8694320
3	Informational	0.8509948
2	Administrative_Duration	0.8483216

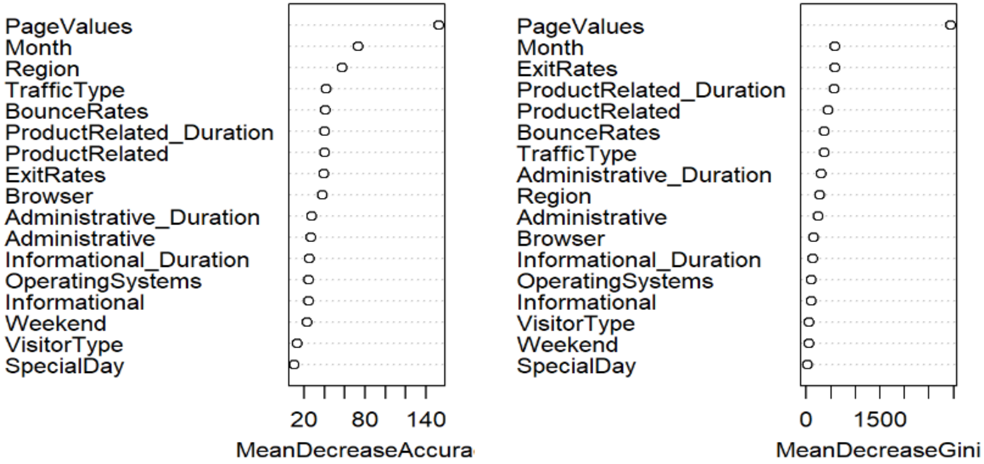
Feature variable importance table

Feature variable importance table	MeanDecreaseAccuracy
Administrative	25.740664
Administrative_Duration	29.428259
Informational	22.772219
Informational_Duration	22.896964
ProductRelated	44.383527
ProductRelated_Duration	37.415946
BounceRates	35.813305
ExitRates	34.568105
PageValues	135.122725
SpecialDay	8.007717
Month	68.915250
OperatingSystems	19.990634
Browser	45.293001
Region	62.972472
TrafficType	39.930089
VisitorType	12.686405
Weekend	25.540213

Number of features vs Accuracy plot



Variable importance plot



Random forest trained on top 10 features

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2913  161
##          1  214  412
##
##                Accuracy : 0.8986
##                  95% CI : (0.8885, 0.9082)
##     No Information Rate : 0.8451
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6269
##
##  Mcnemar's Test P-Value : 0.007247
##
##             Sensitivity : 0.9316
##             Specificity : 0.7190
##          Pos Pred Value : 0.9476
##          Neg Pred Value : 0.6581
##              Prevalence : 0.8451
##          Detection Rate : 0.7873
##    Detection Prevalence : 0.8308
##       Balanced Accuracy : 0.8253
##
##        'Positive' Class : 0
##
```

Top 10 features

PageValues

Month

Region

Browser

ProductRelated

TrafficType

ProductRelated_Duration

BounceRates

ExitRates Administrative_Duration

4) Support Vector Machine

Kernel : Linear Result

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2843  156
##          1  284  417
##
##                Accuracy : 0.8811
##                  95% CI : (0.8702, 0.8913)
##     No Information Rate : 0.8451
##     P-Value [Acc > NIR] : 2.364e-10
##
##                   Kappa : 0.5837
##
##  Mcnemar's Test P-Value : 1.409e-09
##
##             Sensitivity : 0.9092
##             Specificity : 0.7277
##          Pos Pred Value : 0.9480
##          Neg Pred Value : 0.5949
##              Prevalence : 0.8451
##          Detection Rate : 0.7684
##    Detection Prevalence : 0.8105
##       Balanced Accuracy : 0.8185
##
##        'Positive' Class : 0
```

Kernel : Radia Result

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2834  149
##          1  293  424
##
##                Accuracy : 0.8805
##                  95% CI : (0.8697, 0.8908)
##     No Information Rate : 0.8451
##     P-Value [Acc > NIR] : 4.387e-10
##
##                   Kappa : 0.5861
##
##  Mcnemar's Test P-Value : 1.033e-11
##
##             Sensitivity : 0.9063
##             Specificity : 0.7400
##          Pos Pred Value : 0.9501
##          Neg Pred Value : 0.5914
##              Prevalence : 0.8451
##          Detection Rate : 0.7659
##    Detection Prevalence : 0.8062
##       Balanced Accuracy : 0.8231
##
##        'Positive' Class : 0
##
```

5) XG Boost

Training parameters

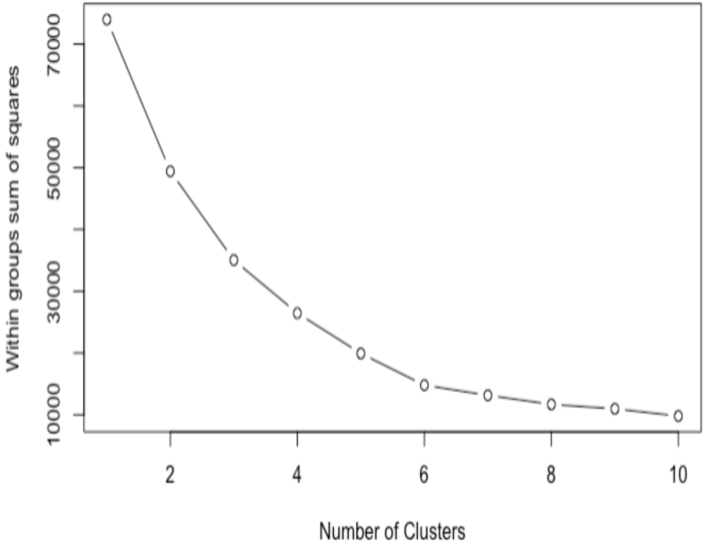
- a) objective = "binary:logistic"
- b) eta = 0.3
- c) max_depth = 6
- d) eval_metric = "auc"
- e) Nrounds = 100
- f) Early stopping rounds = 10
- Stopping. Best iteration: [7]
- train-auc:0.961152
- test-auc:0.927922

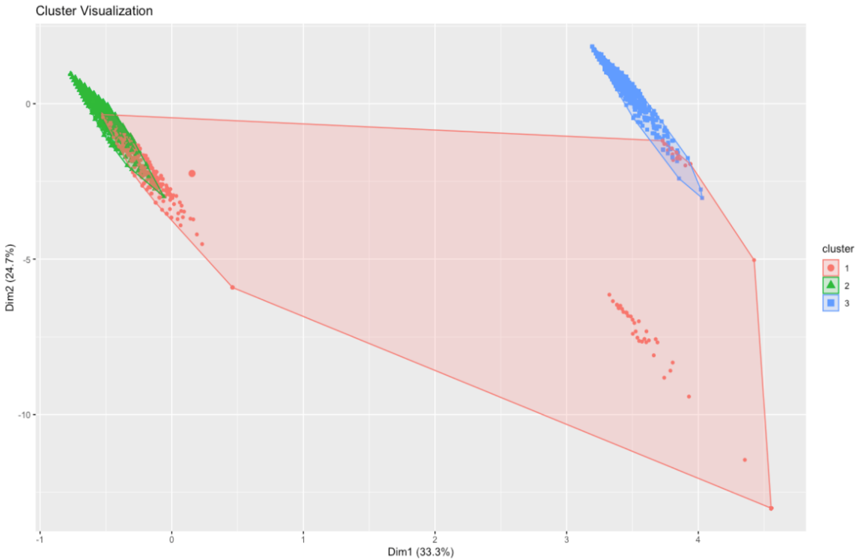
[1] train	auc:0.940757	test	auc:0.909651
[2] train	auc:0.949137	test	auc:0.918012
[3] train	auc:0.952685	test	auc:0.922717
[4] train	auc:0.954565	test	auc:0.926051
[5] train	auc:0.957042	test	auc:0.926611
[6] train	auc:0.958906	test	auc:0.927626
[7] train	auc:0.961152	test	auc:0.927922
[8] train	auc:0.963121	test	auc:0.927406
[9] train	auc:0.963852	test	auc:0.926375
[10] train	auc:0.965559	test	auc:0.925545
[11] train	auc:0.967358	test	auc:0.925664
[12] train	auc:0.969143	test	auc:0.924629
[13] train	auc:0.970182	test	auc:0.924950
[14] train	auc:0.970923	test	auc:0.924509
[15] train	auc:0.973319	test	auc:0.923537
[16] train	auc:0.974383	test	auc:0.923199
[17] train	auc:0.975294	test	auc:0.923111

Un-supervised Learning

K-Means Clustering

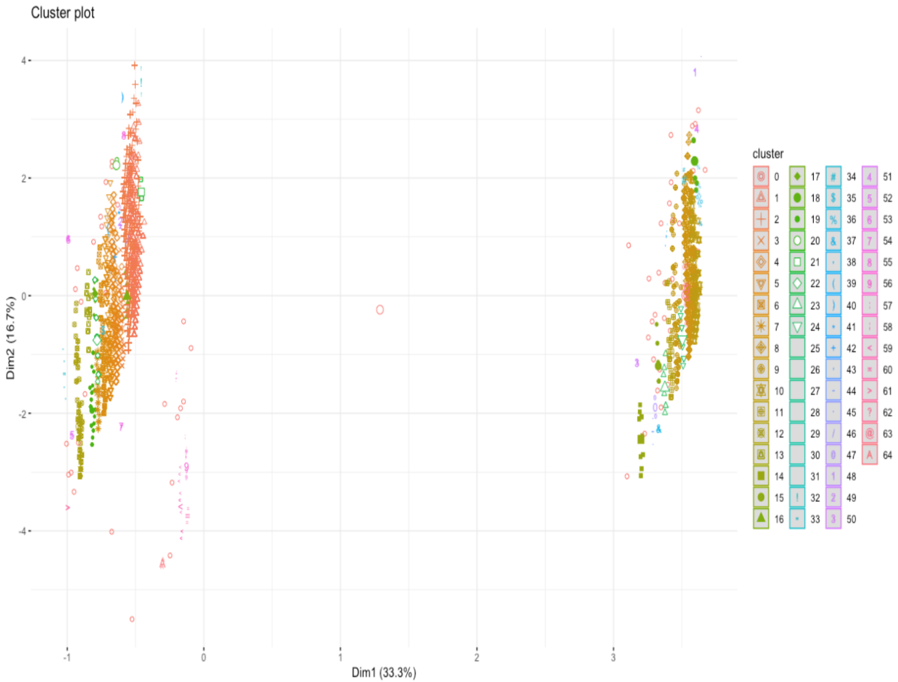
Clusters has been identified using Elbow method, and from the clustered plot we can say that most of the data can be clustered into 3 clusters.





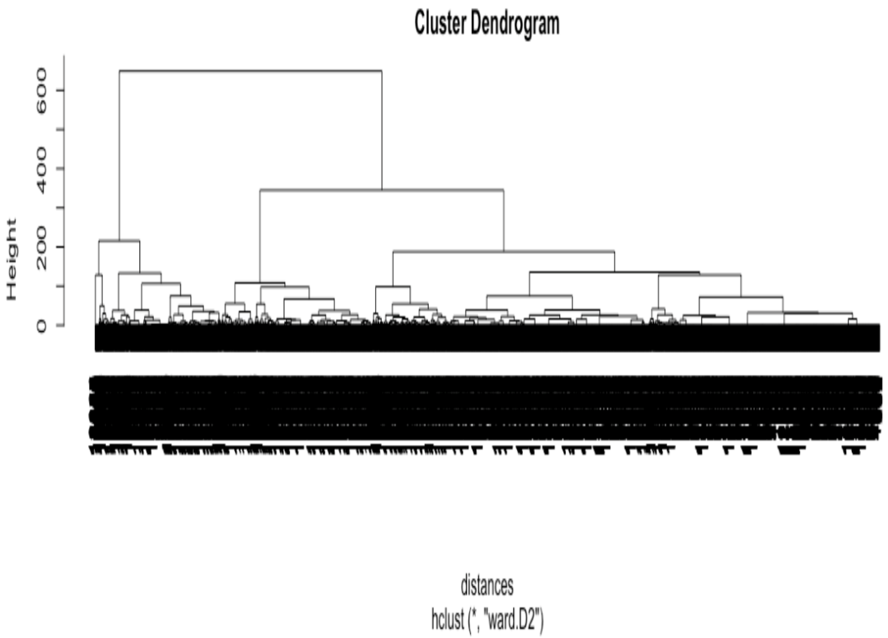
DBScan Clustering

As name says this is a density-based clustering algorithm that groups data points into clusters based on their density. DBSCAN is particularly useful for datasets with irregular shapes and noises. But our dataset was mostly without the noises and hence the results were as below

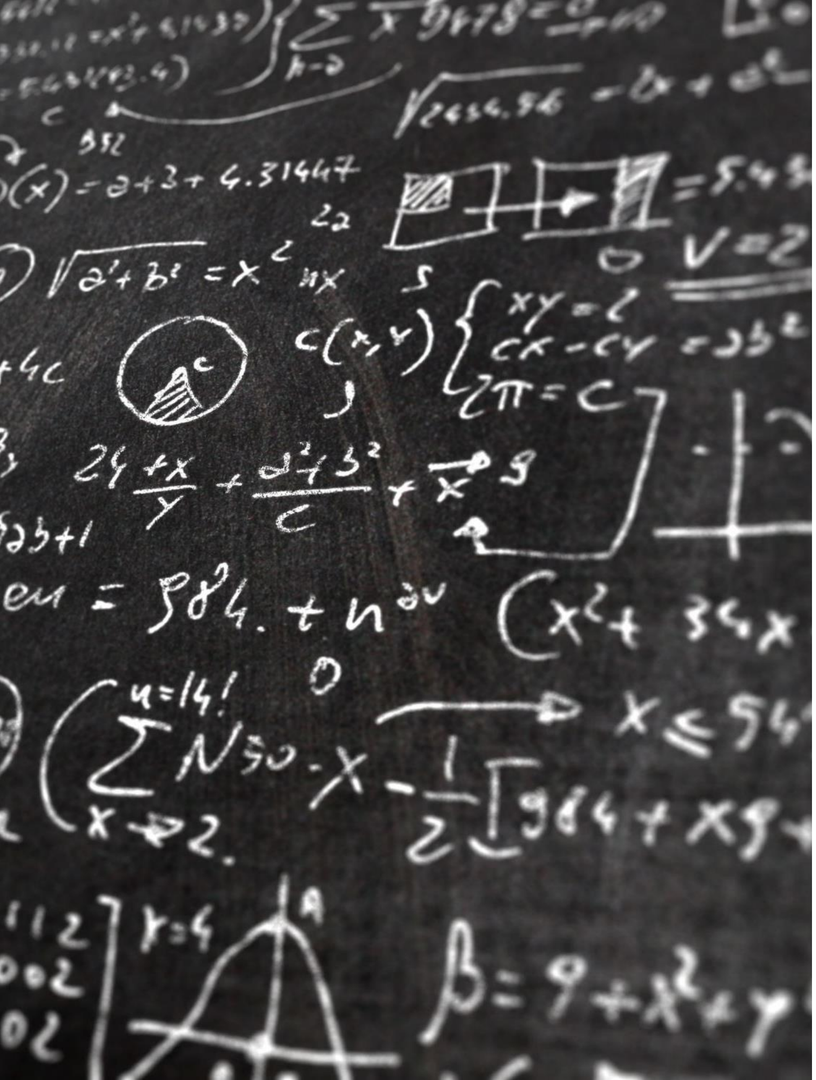


Hierarchical clustering

This is another unsupervised learning algorithm that clusters data points into a tree-like structure based on the similarity between them. Hierarchical clustering can be either agglomerative (bottom-up) or divisive (top-down).



Future Work



Future Work

- We would plan to work more on the data gathering, we did look into it but couldn’t find the similar datasets.
- Work on couple more research question for example, “How does web metrics influence the revenue.”
- Will explore and try to implement MLOps best practices by designing and creating a end-to-end pipelines.
- Would explore Gaussian Mixture Models for the clustering and also analyse the clusters indepth.

Conclusion

Conclusion

- Analyzing number of page visit of 3 different page categories it clearly says that customers are interested more in Product related pages rather than knowing information of the product in detail.
- Revenue is generated by the customers who visit the product page and spend more time on it, which intuitively mean whom ever spends more time on administrative and informational page will only hop around rather than end up buying.
- Discounts can be given to the ones who spend more time on Product related page.
- There is no noticeable disparity in Bounce Rates between customers who made a purchase and those who did not.
- However, customers who ended up making a purchase had lower Exit Rates on average, indicating that they were more likely to remain on the website's pages.
- Additionally, customers who did not make a purchase had significantly lower Page Values, suggesting that they spent less time on related pages.

Bibliography

- [1]. https://jurnalppi.kominfo.go.id/index.php/jppi/article/view/341
- [2]. Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks [https://link.springer.com/article/10.1007/s00521-018-3523-0]
- [3]. Data Clustering: A Review [https://dl.acm.org/doi/pdf/10.1145/331499.331504]
- [4]. A Comparison Study of Credit Card Fraud Detection: Supervised versus Unsupervised [https://arxiv.org/pdf/1904.10604.pdf]
- [5].Real-Time Prediction of Online Shoppers Purchasing Intention Using Random Forest [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7256375/].


THANK
YOU !