2232-INSY-5339-001
PRINCIPLES OF BUSINESS DATA MINING

Spring 2023

FINAL PROJECT REPORT

Submitted by: GROUP - 3
Nilambari Bansode (1002051160)
Nethra Ramesh (1002083168)
Ragavi Kannaiyan (1002076765)
Juanita Jayabalan (1002058143)

Professor: JAYARAJAN SAMUEL

Figures

Executive Summary:

Many people around the world prefer to shop online and buy products from several brands and companies that they cannot find or are not available for purchase in their home countries. Nowadays with the help of new technology and the support of the internet, people from all around the world have started to purchase items online by simply sitting in their homes. For a website, user behavior analysis can be performed which will provide information on the purchase intent of the user based on the website activity.

Project Motivation:

Based on the user's activity within the website we might be able to identify if the user will make a purchase or not. The objective of this project is to answer the question "Which pages need to be targeted and needs improvement, so that the customer experience is improved, and they spend less time loitering around eventually reducing the bounce rate and helping them complete the transaction efficiently."

Data Description:

The dataset has around 18 attributes and 12330 records which have details about the user's activity on a webpage like the number of pages of a specific type that the user visited, the amount of time spent on each of these categories, the percentage of users who enter the website through a specific page and exit without performing any activity, revenue attribute which says whether the user has made a purchase or not, etc. The dataset consists of 10 numerical attributes and 8 categorical attributes.

Number of observations in the dataset	12330
Outcome/Target Variable	Revenue

Categorical Attributes:

Attribute Name	**Attribute Description**	**No. of categorical values**
OperatingSystems	The operating system of the visitor	8
Browser	The browser of the visitor	13
Region	The geographic region from which the session has been started by the visitor	9
TrafficType	Traffic source by which the visitor has arrived at the Web site (e.g., banner, SMS, direct)	20
VisitorType	Visitor type as "New Visitor," "Returning Visitor," and "Other"	3

Weekend	Boolean value indicating whether the date of the visit is weekend	2
Month	Month value of the visit date	12
Revenue	A class label indicating whether the visit has been finalized with a transaction	2

Numerical Attributes:

Attribute name	Attribute description	Max. value
Administrative	Number of pages visited by the visitor about account management	27
Administrative duration	The total amount of time (in seconds) spent by the visitor on account management-related pages	3398
Informational	Number of pages visited by the visitor about the Web site, communication, and address information of the shopping site	24
Informational duration	The total amount of time (in seconds) spent by the visitor on informational pages	2549
Product related	Number of pages visited by visitors about product-related pages	705
Product related duration	The total amount of time (in seconds) spent by the visitor on product-related pages	63,973
Bounce rate	Average bounce rate value of the pages visited by the visitor	0.2
Exit rate	Average exit rate value of the pages visited by the visitor	0.2
Page value	Average page value of the pages visited by the visitor	361
Special day	The closeness of the site visiting time to a special day	1

Data Visualization:

Circle Plot:

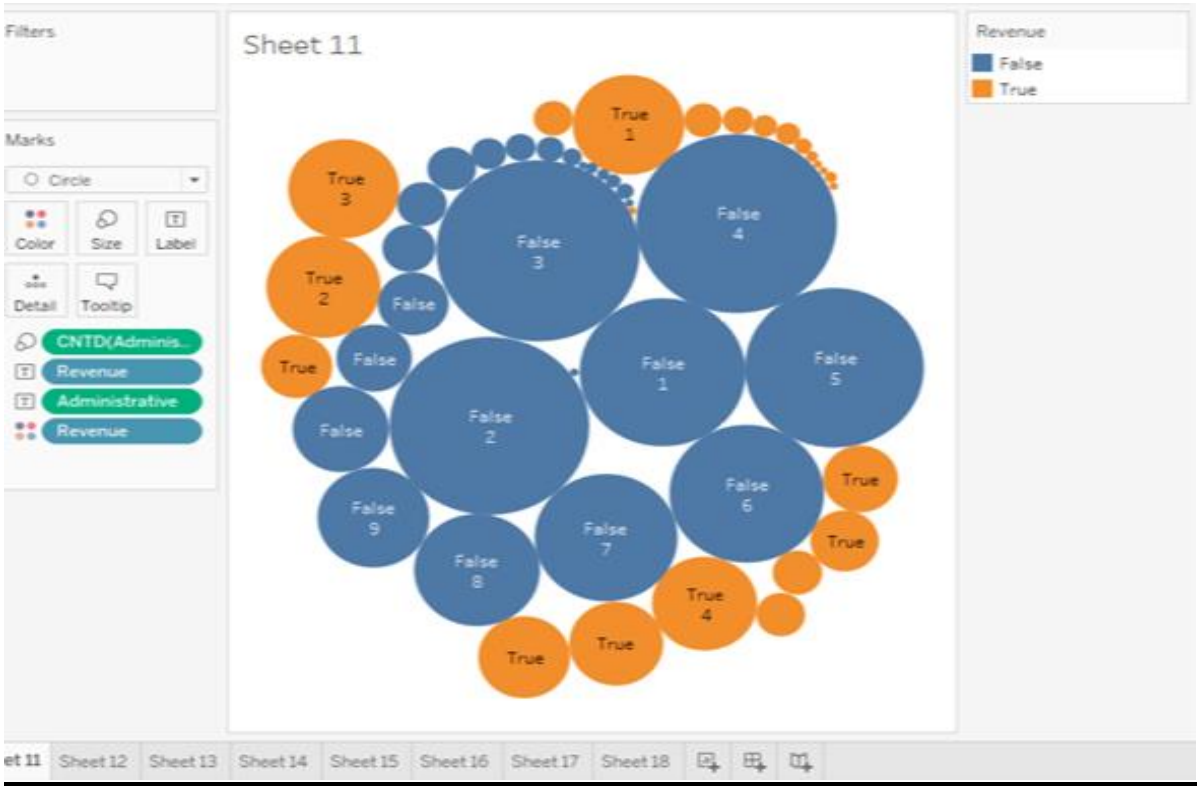


Figure 1

- The circle plot shows the distribution of Revenue (categorical data – True/False) concerning the Administrative and Administrative Duration measures.
- Based on this graph, we do see that the customers who spend more amount of time on the website have not bought the product (False indicates did not buy) and comparatively people who spent less amount of time on the website ended up buying the product.

Bar Graphs:

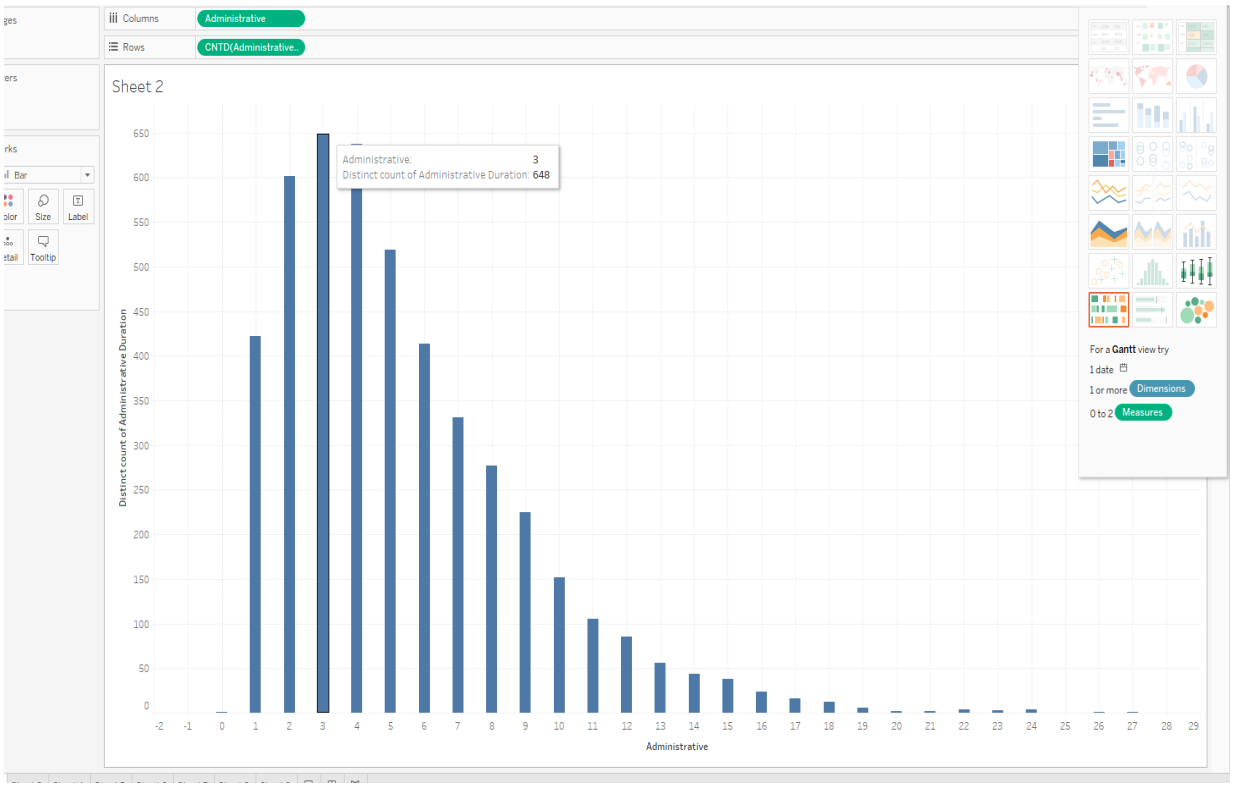


Figure 2

- The bar graph shows the distribution of administrative pages concerning the Administrative Duration measure.
- Based on this graph, we do see that the customers spend more time on administrative page 3 with 648 seconds. And closer to it is page 4 and page 2 too. From this, we could get an idea that these pages need some modification so that the user experience of it is improved.

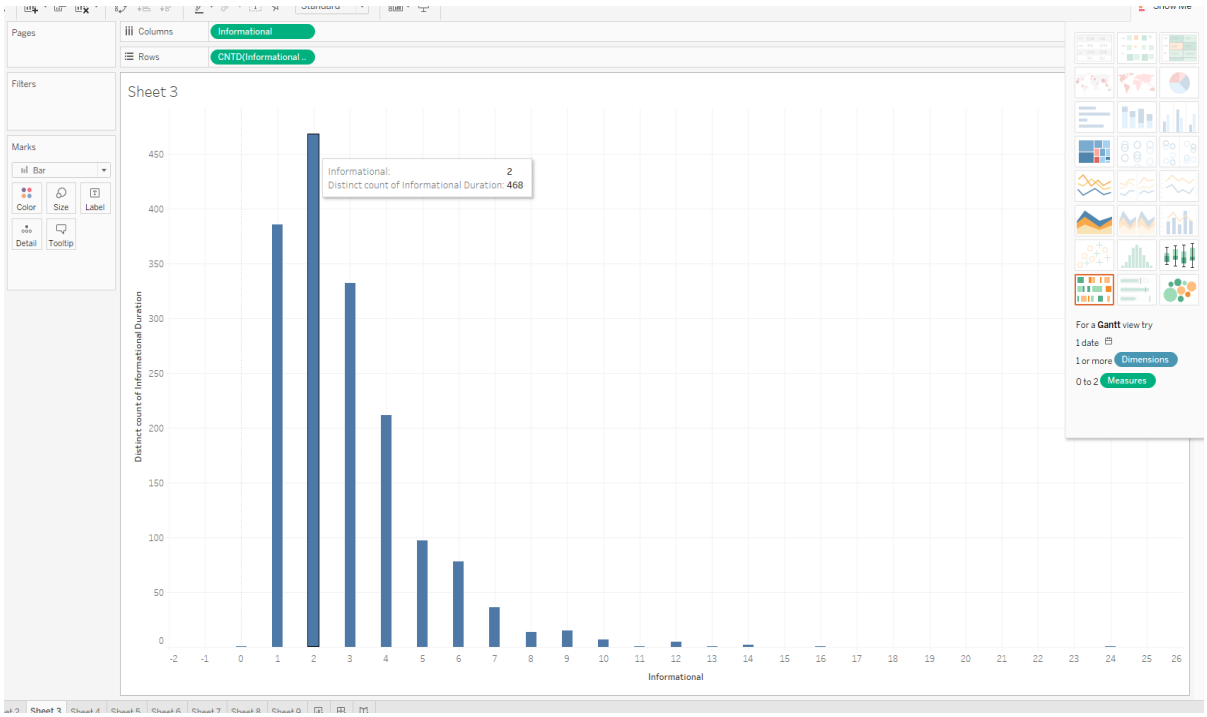


Figure 3

- The bar graph shows the distribution of Informational pages concerning the Informational Duration measure.
- Based on this graph, we do see that the customers spend more time on Informational page 2 with 468 seconds. And closer to it is page 4 and page 1 too. From this, we could get an idea that these pages need some modification so that the user experience of it is improved.

Tree Maps:

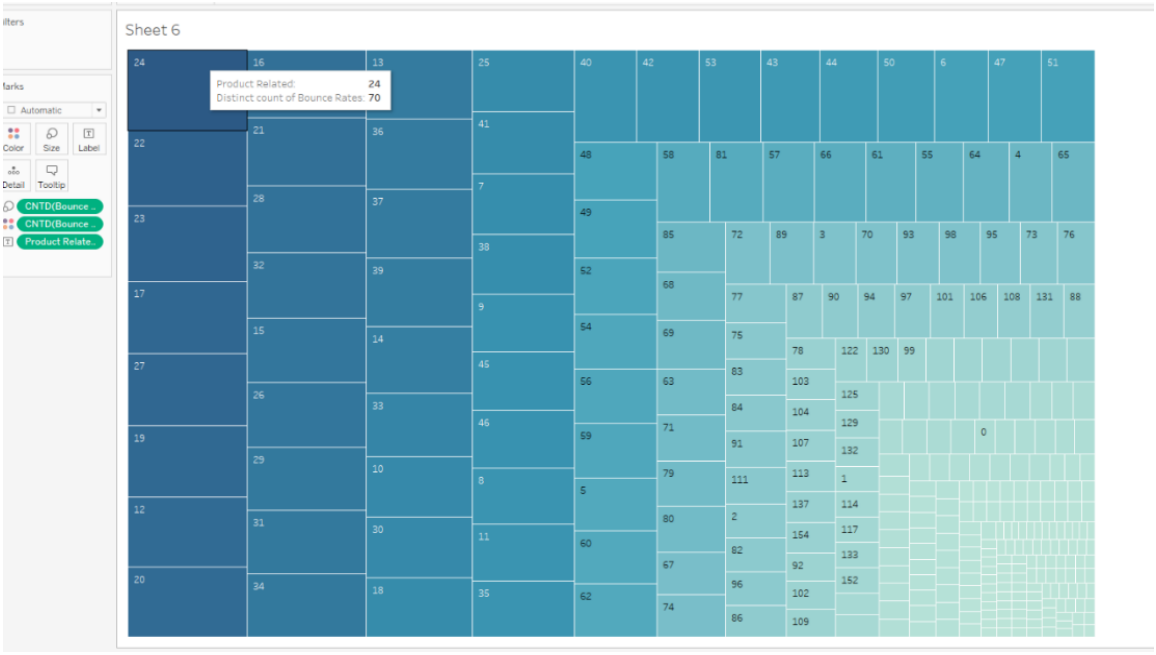


Figure 4

- The above Tree Map shows the plotting of the Product along with Product Duration.
- From this, we can see that page 24 has been spent more time by the customers concerning Product pages.
- The dark blue rectangles show the product-related pages on which longer time was spent and as the color fades the time spent on those respective pages has reduced.

HeatMap:

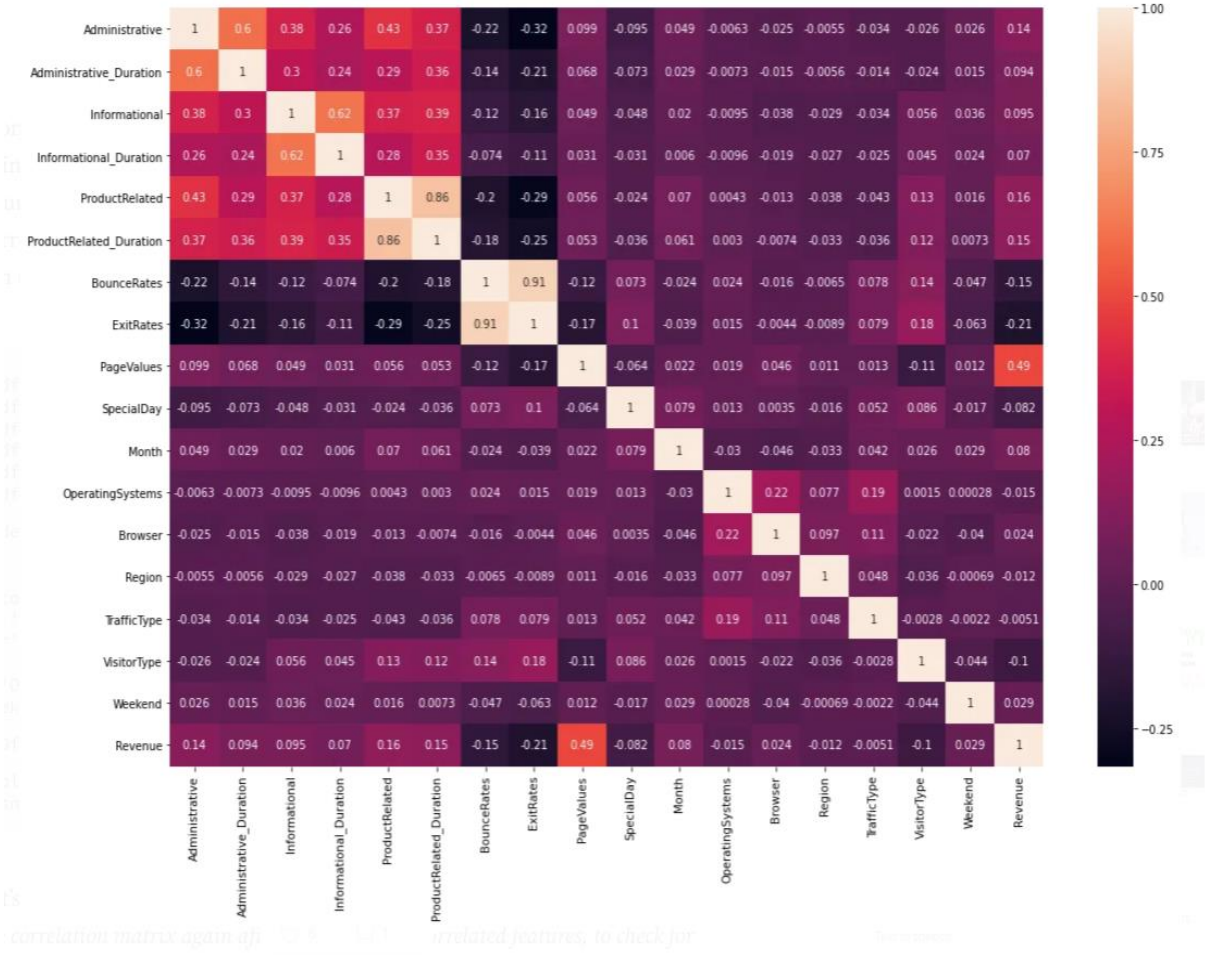


Figure 5

- From the above figure, it is clear that administrative data (both duration and point) are correlated. Information, Product Related, and Rates (Exit and Bounce) have similar Characteristics.
- Page Value seems to have a stronger correlation with Revenue.

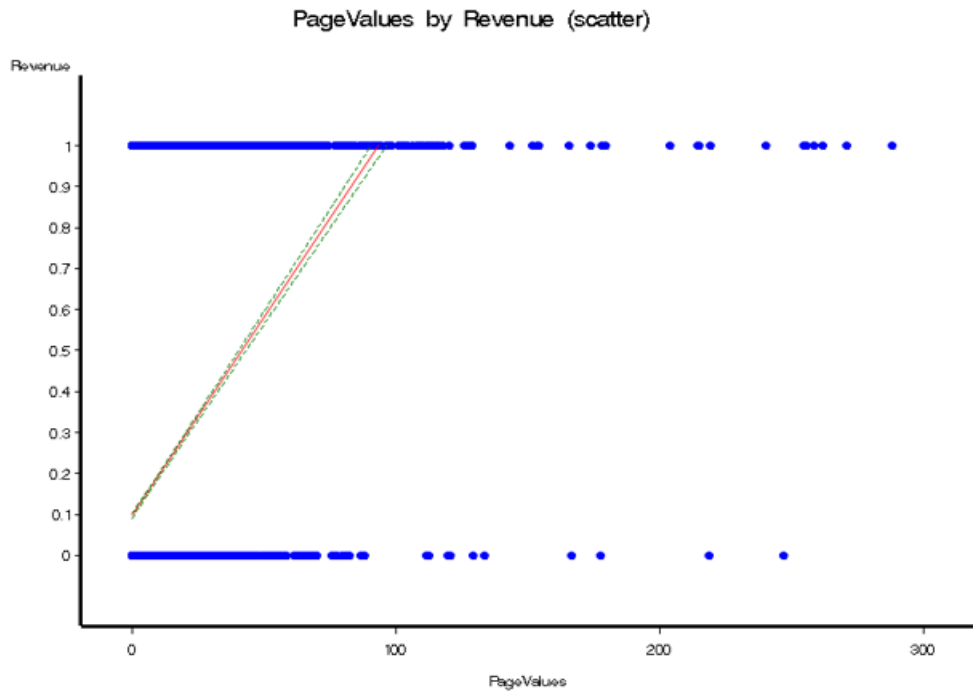


Figure 6

In the above figure, we could see that PageValues have been plotted with Revenue. From the figure we get to know that with the increase in the PageValues, the Revenue also increases.

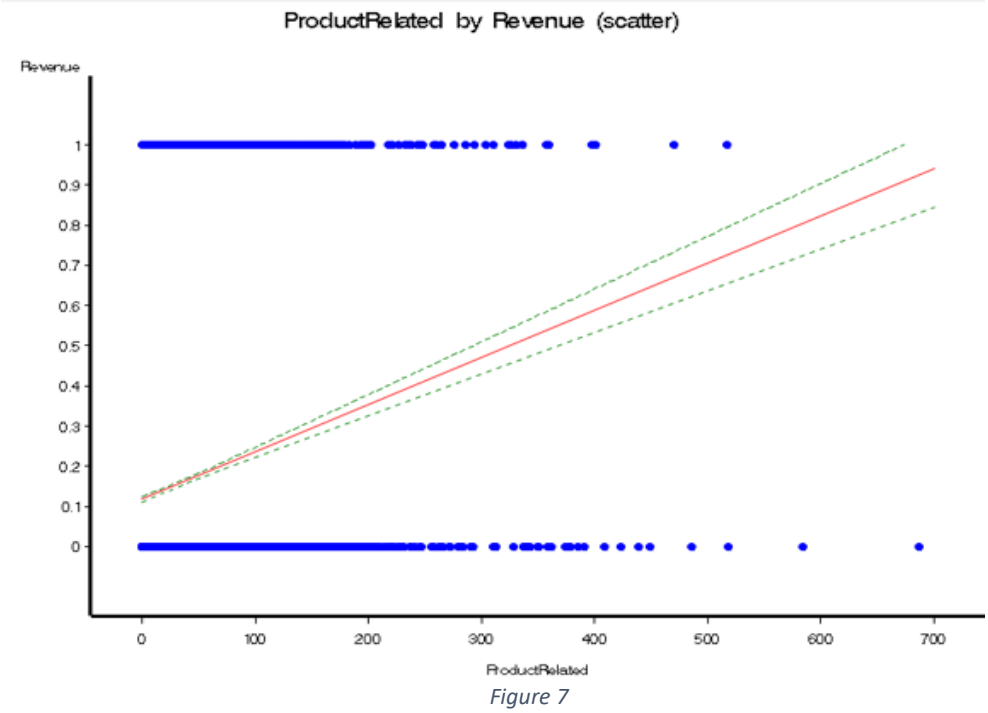


Figure 7

In the above figure, we can see that ProductRelated has been plotted with Revenue. This figure tells us that there is a linear relationship between them.

Data Partitioning:

For this analysis, the dataset with 12,330 records is partitioned into training and testing data for both the regression model and the decision tree. The partition is as below:
Train Data (50%) = 6165
Test Data (50%) = 6165

Ranking of Variables:

Variable Worth is a module in SAS Enterprise Miner which is used to rank the variables based on how it affects the target variable "Revenue".

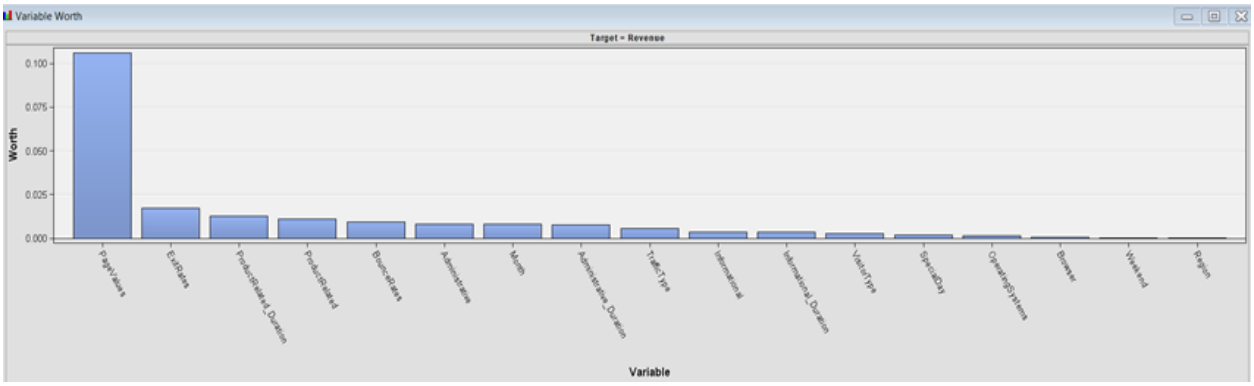


Figure 8

From the above graph, we could see the variable "Page Values" Has a higher effect on the target variable. Other major factors which affected the target variable based on the above chart are "ExitRates", "ProductRelated", "ProductRelated_Duration", and "BouceRates".

Models Used in the Project:

- Decision Tree
- Linear Regression with Forward Selection Model
- Naïve Bayes

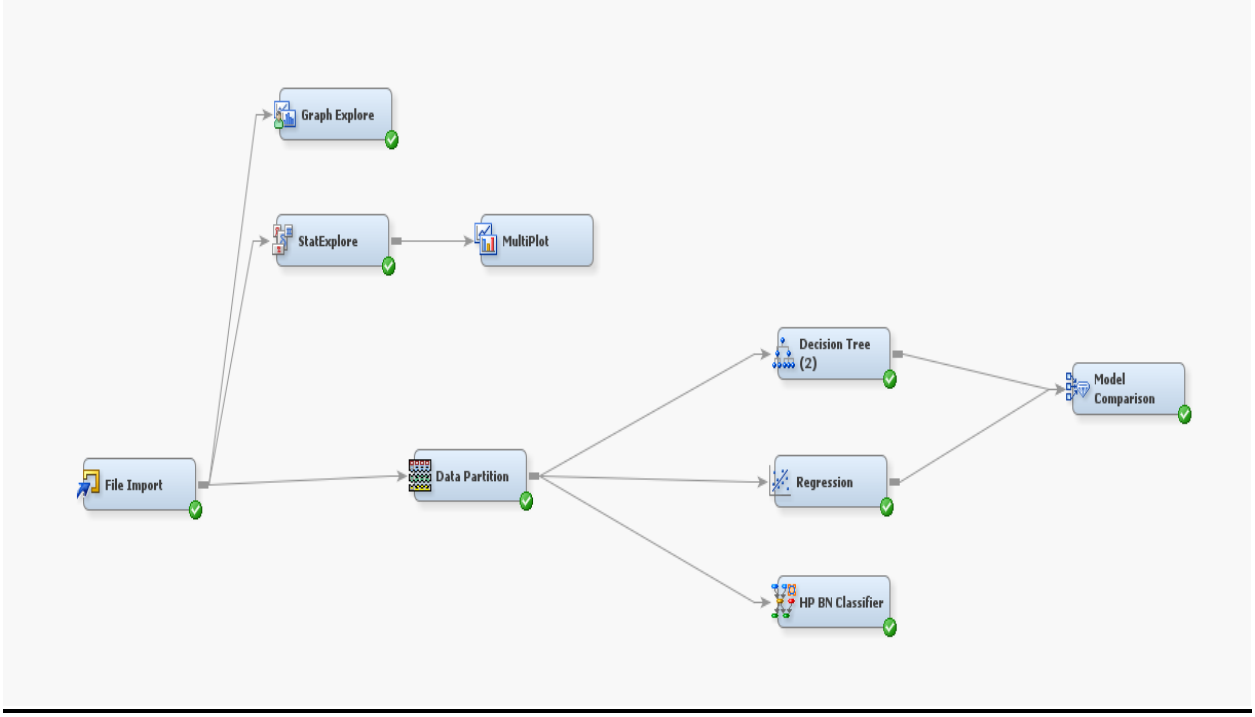


Figure 9

Decision Tree:

- The decision tree is a supervised learning algorithm that partitions variables based on a split criterion.
- Since the dataset we work with is quite versatile, decision trees are an excellent option because they can handle both continuous and categorical variables.

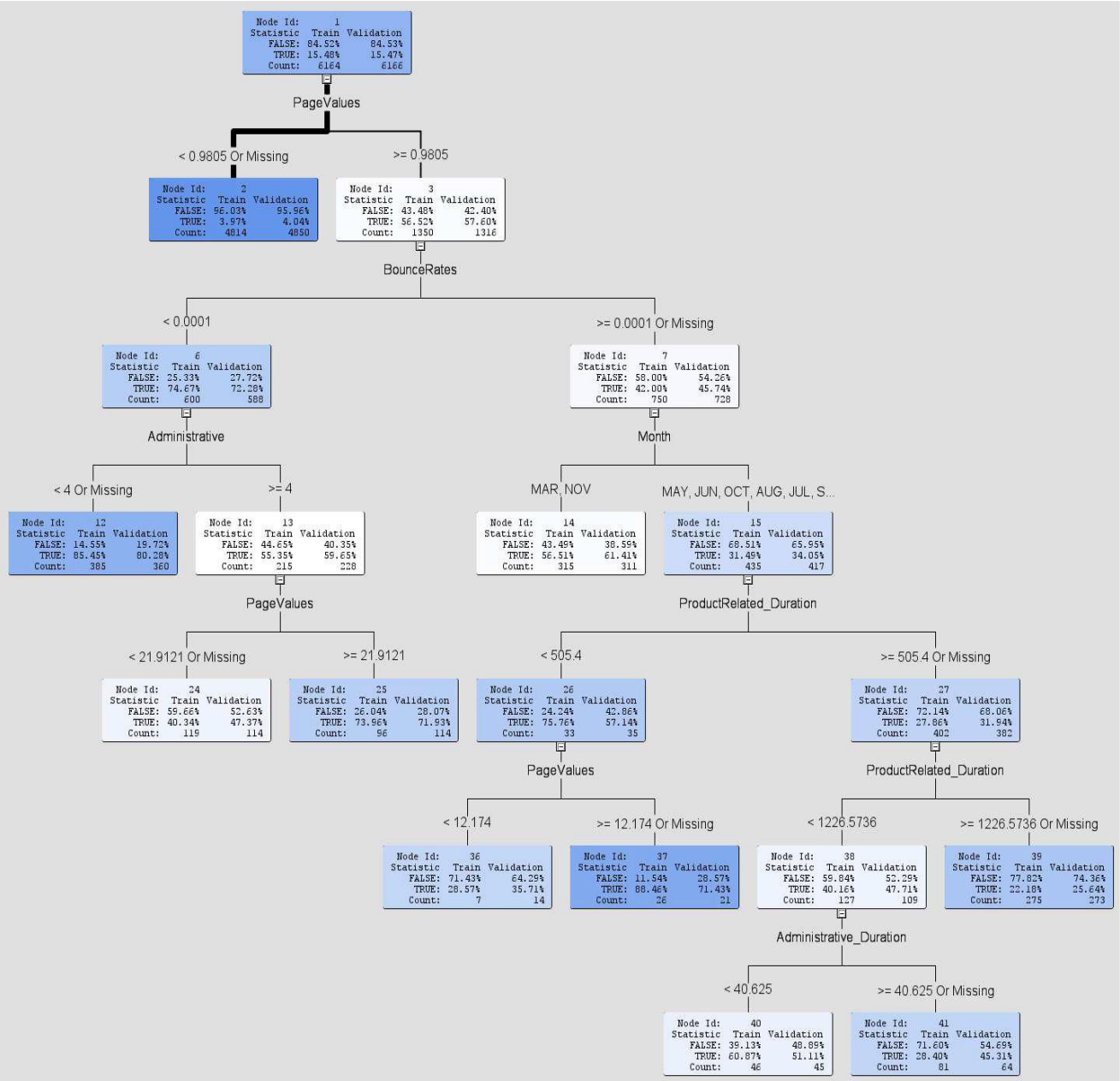


Figure 10

Model Results:

- The decision tree consists of 41 nodes and 10 leaf nodes.
- The first split in the decision tree is based on variable Page values with a cut-off of 0.9805 concerning the target variable Revenue.

Classification Table

Data Role=TRAIN Target Variable=Revenue Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
FALSE	FALSE	93.8633	95.4127	4971	80.6457
TRUE	FALSE	6.1367	34.0671	325	5.2726
FALSE	TRUE	27.5346	4.5873	239	3.8774
TRUE	TRUE	72.4654	65.9329	629	10.2044

Data Role=VALIDATE Target Variable=Revenue Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
FALSE	FALSE	93.3396	95.1842	4961	80.4573
TRUE	FALSE	6.6604	37.1069	354	5.7412
FALSE	TRUE	29.4947	4.8158	251	4.0707
TRUE	TRUE	70.5053	62.8931	600	9.7308

Event Classification Table

Data Role=TRAIN Target=Revenue Target Label=' '

False Negative	True Negative	False Positive	True Positive
325	4971	239	629

Data Role=VALIDATE Target=Revenue Target Label=' '

False Negative	True Negative	False Positive	True Positive
354	4961	251	600

Figure 11

- From the above output we do see that for both the training and test data, the number of true positives and true negatives is significantly greater than the number of false negatives and false positives in both the train and validation sets, indicating good performance of the decision tree model.
- For the training set, we do see that the number of true positives is 629 and the number of true negatives is 4971, which is much higher compared to the false rates with false positives having 239 and false negatives 325.
- A similar kind of behavior is seen with validation sets with the values as follows: True positive – 600, True negative – 4961, False positive – 251, False negative - 354.

Linear Regression – Forward Selection:

- Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables.
- In linear regression the forward selection model starts with no variables in the model followed by testing the addition of each variable using a chosen model fit criterion, adding the variable (if any) whose inclusion gives the most statistically significant improvement of the fit.

Model Results:

- From this model, we do see that the variables ExitRates, PageValues, and ProductRelated_Duration is more statistically significant as compared to other variables in the dataset.

The selected model is the model trained in the last step (Step 7). It consists of the following effects:

Intercept Browser ExitRates Month OperatingSystems PageValues ProductRelated_Duration VisitorType

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
10624.796	7164.118	3460.6776	16	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Browser	1	4.9662	0.0258
ExitRates	1	121.9649	<.0001
Month	9	258.1751	<.0001
OperatingSystems	1	4.0706	0.0436
PageValues	1	1175.5832	<.0001
ProductRelated_Duration	1	59.9604	<.0001

Figure 12

Model Comparison:

- After performing a model comparison between the Decision tree and Linear regression Forward selection we find that the Decision tree performs well.

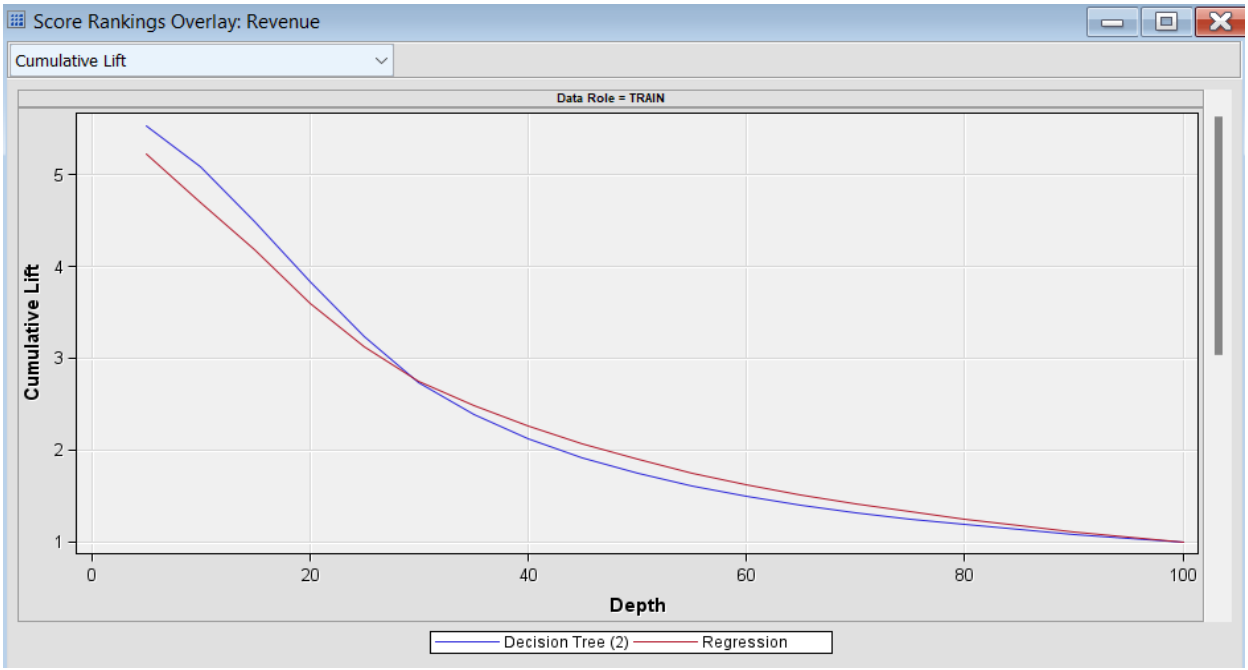
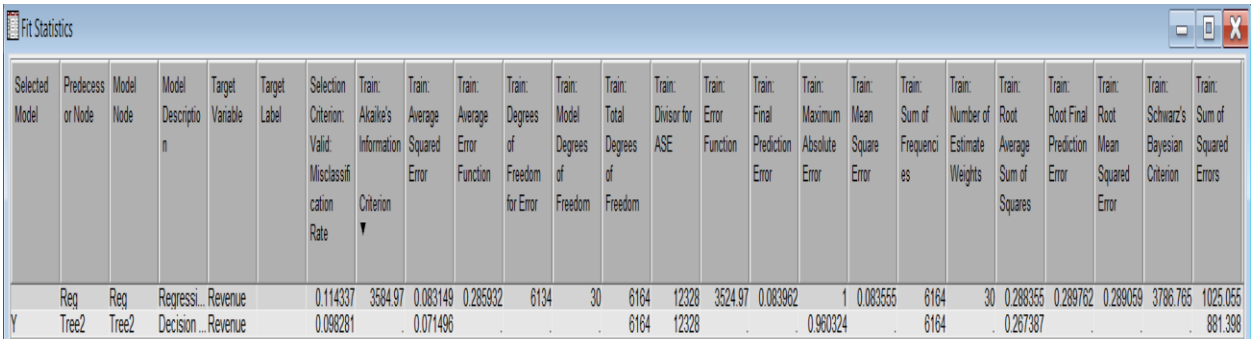


Figure 13



Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Total Degrees of Freedom	Train: Divisor for ASE	Train: Error Function	Train: Final Prediction Error	Train: Maximum Absolute Error	Train: Mean Square Error	Train: Sum of Frequencies	Train: Number of Estimate Weights	Train: Root Average Sum of Squares	Train: Root Final Prediction Error	Train: Root Mean Squared Error	Train: Schwarz's Bayesian Criterion	Train: Sum of Squared Errors
	Reg	Reg	Regressi...	Revenue		0.114337	3584.97	0.083149	0.285932	6134	30	6164	12328	3524.97	0.083962	1	0.083555	6164	30	0.288355	0.289762	0.289059	3786.766	1025.055
Y	Tree2	Tree2	Decision ...	Revenue		0.098281	.	0.071496	.	.	.	6164	12328	.	.	0.960324	.	6164	.	0.267387	.	.	.	881.398

Figure 14

- The above output shows that the decision tree model has a smaller misclassification rate (represents the proportion of misclassified instances or observations in the total number of instances or observations) in comparison to the linear regression.

Analysis to find the pages which need to be improved:

After performing an analysis of both the models we do see that the significant page-related variable is Product related. As our main objective is to find the specific pages which need to be improved, we will run the Naïve Bayes model for the product-related pages to find the probabilities of the pages.

Naïve Bayes:

Using Naïve Bayes, we predicted the probability of Product related pages whose user experience can be improved as follows:

- Segregation of Product related pages into quantiles and setting threshold value as 2266 here as 75% of people are ending up buying products within 2266 seconds and below that. The cut-off value can even be 1109 depending on the number of pages that businesspeople want to improve.

- We created an additional variable of ProductRelated_Duration_quality with values Good & Bad

  Good – Customers buy a product with 2266 Seconds and less than that.
  Bad – Customers buying a product with more than 2266 Seconds.

- Based on this new variable we ran a condition using Naïve Bayes that if a customer spends a good amount of time on the product-related page and still does not end up buying the product which sets revenue as equal to false then the pages along with their probabilities would be displayed
- We picked the pages with probability 70% and above and the table of pages along with the probability is shown below.

```
ds.groupby(["Revenue"])["ProductRelated_Duration"].quantile([.50,.75,.90,.95])

Revenue
False    0.50     510.190000
         0.75    1331.816667
         0.90    2632.117679
         0.95    3886.583454
True     0.50    1109.906250
         0.75    2266.011310
         0.90    4532.587520
         0.95    6225.247539
Name: ProductRelated_Duration, dtype: float64
```

Figure 15

After running Naïve Bayes with the cut-off value, we found the list of pages with the greater probability which can be improvised.

Page	Probability
22	0.819360599
30	0.790977066
32	0.774968792
21	0.771915109
10	0.760416332
19	0.747459845
27	0.732223896
13	0.716642759
20	0.705679711

Conclusion

Pages with reduced bounce rates have been demonstrated to be a reliable predictor of whether a visit results in sales.

- The likelihood that a customer would make a purchase rises by 55% for every 0.04 decrease in bounce rates.
- Additional insights can be gained by examining the characteristics of pages with lower bounce rates.

A higher page value is a reliable sign that a visit will result in sales.

- Pages with a Page Value score greater than 0 receive 43 percent of all visits and earn **income**.
- The probability of a client making a purchase rises by 8 times for every 19 units increase in Page Value.

The best traffic ratios are generated in November, however, months like December also have a lot of traffic.

- When compared to the previous months, November's chances of making a purchase rose by 275%.
- To increase revenue, December traffic might be targeted for conversion.

Suggestions for Business Improvement

- From the Regression model, we got to know that Product Related pages have more significance on Revenue.
- Based on our results of Naïve Bayes, we conclude on specific Product Related Pages which need improvement in their User Interface as the time spent on them is Good but still the customers do not buy the product.
- As a recommendation, we would advise the Manager to improve the user experience of those pages, which is probably higher than 70%.
- In the future, if we have a larger budget, we can even reduce the threshold of 70% to 60% or 50% which will in turn give us a greater number of pages whose user experience needs to be improved so that the customer can end up buying the product efficiently.

Resources:

- SAS Enterprise Miner 15.2
- Tableau Desktop 2021.2.1
- Microsoft Excel
- Microsoft Word
- Python
- https://www.kaggle.com/

Datasets:

https://www.kaggle.com/datasets/imakash3011/online-shoppers-purchasing-intention-dataset

online_shoppers_intention.csv