**PREDICTION OF ONLINE SHOPPERS PURCHASING INTENTION MODEL**

Diagram

Description automatically generated

**Contents:**

1. Executive Summary

1.1. Project Motivation

2. Data Source

2.1. Source Link

2.2. Data Description

3. Data Preprocessing & Exploratory Data Analysis

3.1. Count of Revenue Generated

3.2. Conversion Rate Over the Year

3.3. Page Value Over the Year

3.4. Type of Visitors

4. BI Modelling

4.1. Model 1 – Decision Tree

4.2. Model 2 – Logistic Regression

4.3. Model 3 - Neural Network’s

5. Conclusion

6. References

**1. EXECUTIVE SUMMARY:**

The E-Commerce industry is one of the world's major industries that must be constantly updated with cutting-edge technology to provide the best services to customers, with the goal of predicting online shoppers' purchasing decisions. Many people who visit ecommerce websites may not intend to buy anything. This could be because of several factors. However, we can determine whether a user is likely to purchase or not based on their activity on the ecommerce website. We used Google Analytics data from an ecommerce website to investigate the possibility of predicting customer purchase intent in this project. Machine Learning algorithms are used to create highly accurate prediction models. Ecommerce businesses can benefit greatly from the ability to predict customer purchase intent because it allows them to better understand the digital retail space.

* 1. **PROJECT MOTIVATION:**

Shopping dynamics are changing around the world as retail shopping continues to shift to E-commerce shopping. E-commerce is already a significant retail market. Customers frequently browse e-commerce site pages before placing orders or abandoning their browsing without making a purchase. This information can help businesses better cater to customer preferences and mutually benefit both the business and the customers by recommending products tailored to each customer and thus increasing sales for the businesses.

1. **DATA SOURCE:**

The dataset is open-source, and it is available to the public on the website UCI Machine Learning Repository.

**2.1 SOURCE LINK:**

<https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset>

**Number of Records:** 12330

**Number of Columns:** 18

**Size of the dataset:** 1047 KB

**2.2. DATA DESCRIPTION:**

The description of each column is shown below.

|  |  |  |
| --- | --- | --- |
| **S.NO** | **VARIABLES** | **DISCRIPTION** |
| 1 | Administrative | Represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories. |
| 2 | Administrative Duration | Represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories. |
| 3 | Informational | Represents the detailed information’ regarding products. |
| 4 | Informational Duration | Represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories. |
| 5 | Product Related | Represent the different types of product details. |
| 6 | Product Related Duration | Represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories. |
| 7 | Bounce Rate | Feature for a web page refers to the percentage of visitors who enter the site from that page and then leave without triggering any other requests to the analytics server during that session. |
| 8 | Exit Rate | Feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session. |
| 9 | Page Value | Feature represents the average value for a web page that a user visited before completing an e-commerce transaction. |
| 10 | Special Day | Feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction. |
| 11 | Month | Value indicating whether the date of the visit is month of the year. |
| 12 | Operating Systems | A Boolean value indicating whether the date of the visit. |
| 13 | Browser | The Number of types of Customers visited. |
| 14 | Region | A Boolean value indicating whether the date of the visit is weekend, and month of the year. |
| 15 | Traffic Type | The Number of types of Customers visited. |
| 16 | Visitor Type | Feature indicates whether the visitor is a new visitor or returning. |
| 17 | Weekend | Indicates whether the day of the week that the session started falls on the weekend or not. |
| 18 | Revenue | The target “Revenue” demonstrates that the majority of customers failed to complete the purchasing |

1. **DATA PREPROCESSING & EXPLORATORY DATA ANALYSIS:**
   1. **Correlation Plot (Before Modification)**

Chart, bar chart

Description automatically generated

* 1. **Correlation Plot (After Modification -cutoff 0.60)**

Chart, scatter chart

Description automatically generated

Our dataset had 0 null values and based on Correlation Matrix, we analyzed that there were a few pair of numerical columns that were highly correlated to each other. To avoid the impact of their correlation on our model’s accuracy we decided to merge each of the pair of correlated columns into single column as a new attribute. We now have a total of 14 columns on which we are doing our model building. We also get an indication that the variable PageValues has potentially high correlation with our target output variable.

* 1. **Count of Revenue Generated:**

Chart

Description automatically generated

* 1. **Conversion Rate Over the Year:**

Chart

Description automatically generated

* 1. **Page Value Over the Year:**

Chart, line chart

Description automatically generated

* 1. **Types of Visitors:**

Chart, pie chart

Description automatically generated

1. **BI MODELLING:**
   1. **Model 1 – Decision Tree:**

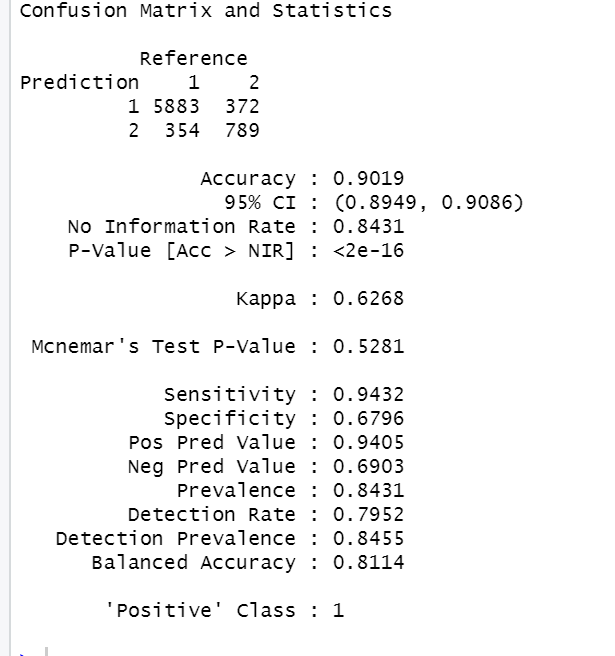
The Decision Tree is a powerful and widely used tool for classification and prediction. A flowchart is similar to a tree structure, with each internal node representing a test on an attribute, each branch representing a test outcome, and each leaf node (terminal node) holding a class label.

**DEFAULT TREE:**

Diagram

Description automatically generated

**Model Accuracy of Training Data of the Default Tree – 90.19%:**



**Model Accuracy of Validation Data of the Default Tree – 89.29%:**

A screenshot of a computer

Description automatically generated with medium confidence

**Deepest Tree:**

To the point where misclassification rate for training dataset is 0%

A picture containing letter

Description automatically generated

**Model Accuracy of Training Data of the Deepest Tree – 100%:**

A screenshot of a computer

Description automatically generated with low confidence

**Model Accuracy of Validation Data of the Deepest Tree- 86.64:**

A screenshot of a computer

Description automatically generated with low confidence

Thus, we can see while the training data shows 100% accuracy the validation data for the deepest tree is

only at 86.64%. This suggests overfitting problem and that we need to prune the tree to get desired level of accuracy in our model.

**Post-Pruning:**

Finding the point where misclassification rate in the validation dataset is the minimum.

Conclusion at nsplit 9 our tree will show the lowest misclassification rate in the validation dataset

as per cp table**.**

Text, letter

Description automatically generated

Table

Description automatically generated

Text

Description automatically generated

**Pruned Tree:**

Diagram

Description automatically generated

**Model Accuracy of Training Data of the Pruned Tree – 90.62%:**

A screenshot of a computer

Description automatically generated with low confidence

**Model Accuracy of Validation Data of the Pruned Tree – 89.31%:**

A screenshot of a computer

Description automatically generated with medium confidence

**Accuracy Comparison:**

We are getting highest accuracy level with the pruned tree at 89.31%.

Table

Description automatically generated

**Applying Random Forest:**

Not Intuitive Anymore as we have 500 trees, we can explain decision making/ data-

driven insights

Chart, line chart

Description automatically generated

**Model Accuracy for Training Data-97.55%:**

A screenshot of a computer

Description automatically generated with medium confidence

**Model Accuracy for Validation Data-89.66%:**

A screenshot of a computer

Description automatically generated with low confidence

**Applying Boosted Trees:**

**Models Accuracy for Training Data-91.89%:**

A screenshot of a computer

Description automatically generated with low confidence

**Model Accuracy for Validation Data -89.56%:**

A screenshot of a computer

Description automatically generated with low confidence

**Final Accuracy Comparison Table:**

Table

Description automatically generated

**Inference:**

* Although, the final accuracy for Boosted Trees is slightly higher than the Pruned Tree we are going to go ahead with the pruned tree as it has a lower difference between training and validation dataset accuracy.
* The variables mentioned above are the most important feature of the dataset for the pruned tree and default tree. Four input variables (PageValues, Bounce\_by\_exit, Month, Adim\_per\_dur) are coming out as more important for output variable revenue as also suggested by Logistic Regression Model.
* Comparing the Sensitivity and Specificity of the three models (default tree, deepest tree, pruned tree, Applying Random Forest, and Applying Boosted Tree).Sensitivity is the metric evaluates a model’s ability to predict true positive of each available category. Specificity is the metric evaluates a model’s ability to predict true negatives of each available category.

**Sensitivity = True Positives/ True Positives + False Negatives**

**Specificity = True Negatives/ True Negatives + False Positives**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Default Tree | Deepest Tree | Pruned Tree | Random Forest | Boosted Tree |
| Sensitivity | 0.9374 | 0.9195 | 0.9450 | 0.9565 | 0.9496 |
| Specificity | 0.6439 | 0.5689 | 0.6024 | 0.5609 | 0.5930 |

* 1. **Model 2 – Logistic Regression:**

The relationship between the dependent variable and one or more independent variables can be better understood using logistic regression. It is used to predict Customer Purchase accuracy rate for revenue when the dependent variable (target) is categorical.

**General Logistic Regression:**

Text, table

Description automatically generated

Text

Description automatically generated

**First 5 Actual and Predicted Records:**

Graphical user interface, text, application

Description automatically generated

**Model Accuracy of validation data actual and predicted records-88.4%:**

Table

Description automatically generated with medium confidence

Text

Description automatically generated

**Model Selection for Full Logistic Regression:**

Text

Description automatically generated

|  |  |
| --- | --- |
| Text  Description automatically generated with medium confidence | Table  Description automatically generated with medium confidence |
| Table  Description automatically generated | Text  Description automatically generated |

Text

Description automatically generated

Text

Description automatically generated with medium confidence

**Model Accuracy of Validation data backward full logistic regression-88.36%:**

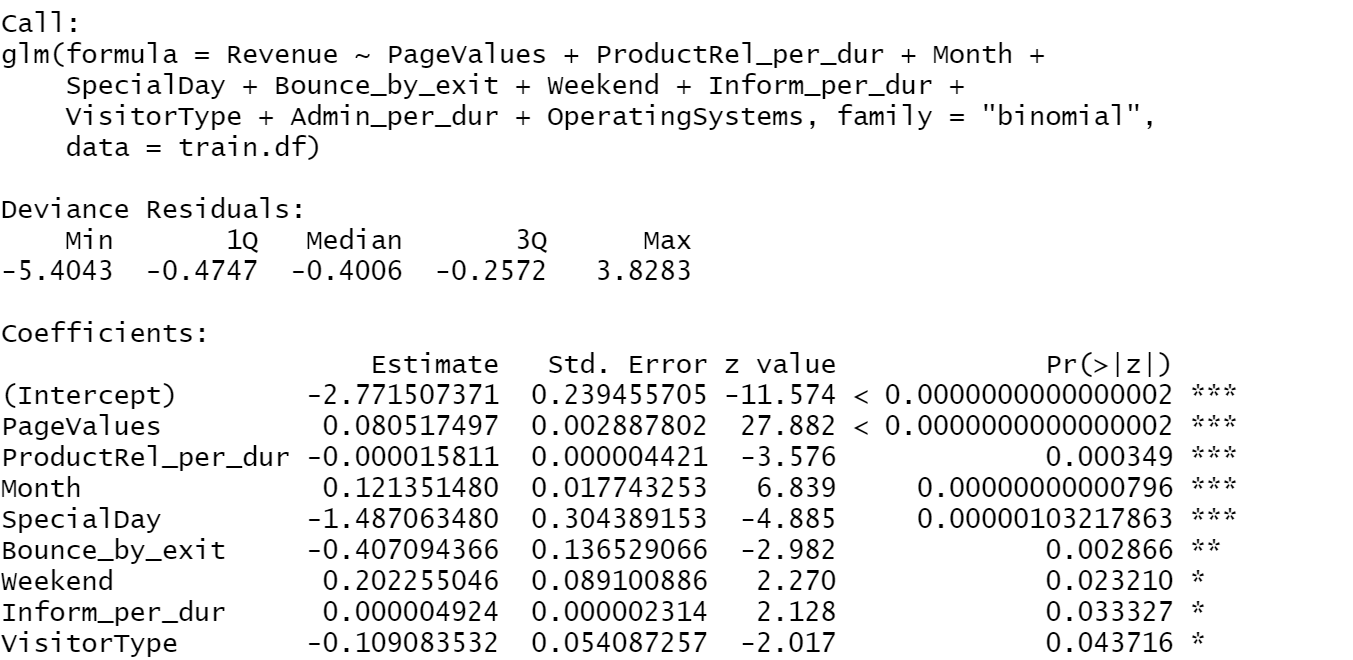
Table

Description automatically generated

A picture containing table

Description automatically generated

**Final General Logistic Regression:**



Text

Description automatically generated

**Model Accuracy of Validation data full logistic regression-88.36%:**

Text

Description automatically generated with low confidence

Text

Description automatically generated

**Inference:**

* We can infer that the odds of increasing the satisfaction level are more inclined towards the variables PageValues, Bounce\_by\_exit, Month, Adim\_per\_dur) which again agrees with the output suggested by Decision Tree analysis done above.
* Comparing the testing accuracy of the two models (backward and full logistic regression), the validation accuracy of both models is same 88.36%.
* Comparing Logistic Regression and Decision Tree, Decision Tree is better fit for this Model

**4.3 Model 3 – Neural Network’s:**

The neural network uses parallel information processing to extract meaningful information and

detect hidden patterns in complex data sets.

A picture containing radar chart

Description automatically generated

A picture containing map

Description automatically generated

Diagram

Description automatically generated with low confidence

A picture containing map

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated with medium confidence

**Inferences:**

* The Accuracy of validation dataset for neural networks is 88.36%

1. **CONCLUSION:**

* **We recommend Decision Tree for this dataset after comparing all the results for three models.** Though the scores are not the best, they are influenced by the dataset's extreme outliers and skewness. As a result, resampling the data or adding more data will affect the model's accuracy and may improve predictions.
* **As highlighted by the models the dependent variables** Page Values, Bounce\_by\_exit, Month, Adim\_per\_dur) are likely to have more impact on the output variable revenue so we could use them to segment our audience for marketing campaigns and get more desirable marketing outcomes.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Decision Tree** | **Logistic Regression** | **Neural Network’s** |
| **Validation Dataset** | 89.31% | 88.36% | 81.39% |

**Insights from EDA:**

* Holiday season October and November witnessed comparatively higher conversion rate than other months with highest being in November.
* However, we find the month of May to be particularly interesting as it has the highest number of visits but below average conversion rate of 11%. Average conversion rate is 14.4%. Need to get further data to bifurcate the visitor base and investigate the reasons for this.
* Almost 86% of the visitors were loyal/returning visitors we need to work on targeting them in the right month and the right time and tapping into their full potential to contribute towards revenue. **We can use this model to run a targeted Loyalty Program to retain these returning customers.**

**Potential Economic Impact:**

* Let’s assume the company would plan to run a promotional (Price drop) mailing campaign targeted towards those customers that are predicted to generate revenue while visiting.
* Average Cost of mailing $1 and average revenue from respondent $50.
* If we follow our decision tree model and the final confusion matrix to find the target audience we can see we would be able to generate net profit of $176,490 by running the campaign.
* This is a very simple model to explain the impact of targeting using appropriate techniques and algorithms.

Table

Description automatically generated

1. **REFERENCES:**

* <https://www.pluralsight.com/guides/explore-r-libraries:-rpart>
* <https://thatascience.com/learn-machine-learning/gini-entropy/#:~:text=Gini%20index%20and%20entropy%20is,only%20one%20class%20is%20pure>.
* <https://discuss.analyticsvidhya.com/t/how-does-complexity-parameter-cp-work-in-decision-tree/6589>
* <https://stats.stackexchange.com/questions/524510/logistic-regression-what-is-the-link-between-the-binomial-family-and-the-binomi>
* <https://www.geeksforgeeks.org/get-or-set-dimensions-of-a-matrix-in-r-programming-dim-function/#:~:text=dim()%20function%20in%20R,matrix%2C%20array%20or%20data%20frame.&text=Parameters%3A,array%2C%20matrix%20or%20data%20frame>.
* <https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/sample>
* <https://stackoverflow.com/questions/34617883/how-to-remove-multiple-columns-in-r-dataframe>
* https://machinelearningmastery.com/confusion-matrix-machine-learning/