Analyzing online shopping behavior with the help of feature vectors containing web session information.

Yash Pandey
Dept Name: Data Science
State University of New York at Buffalo
Buffalo, USA
yashpand@buffalo.edu

Zaid Chaudhary
Dept Name: Data Science
State University of New York at Buffalo
Buffalo,USA
mohdzaid@buffalo.edu

Piyoosh Kumar
Dept Name: Data Science
State University of New York at Buffalo
Buffalo, USA
piyooshk@buffalo.edu

Naresh Yadav Pochaveni
Dept Name: Data Science
State University of New York at Buffalo
Buffalo, USA
nareshya@buffalo.edu

***Abstract*— In this paper, we have performed an analysis on the shoppers on the e-commerce websites using their session data available to predict whether a person surfing through the ecommerce website will buy a product from their website or not. This finding can help the e-commerce companies to show relevant content to the users who show positive purchasing intent. This can help them gain profit by attracting more customers and boosting their sales. We have performed analysis using two classification models and will be working with more in coming future. The major goal was to identify important variables that contribute the most to predicting shopper behavior and to provide prioritized critical suggestions and performance improvements based on those measures. Revenue is the interest property that indicates whether a purchase was made.**

Keywords- e-commerce, analysis, shoppers, purchasing intent, classification

I. Introduction

The expansion in e-commerce usage over the last several years has produced market potential, but the fact that conversion rates have not improved at the same rate necessitates the development of solutions that offer online buyers personalized promotions. This has a significant impact on time management, buy conversion rates, and sales numbers. Many e-commerce businesses invest in early detection and behavioral prediction technologies to get insights and boost revenue by keeping customers engaged with their products.

A real-time online shopper behavior analysis system is proposed in this paper. In real-time during a session, the suggested system assigns a score to the visitor's purchasing intent. We assess the performance of different machine learning algorithms under various scenarios using data from an online shop.

Our outcome assists businesses in detecting blind spots to improve conversion rates. It also enables the businesses to make specific adjustments to the website. It enables to provide consumers with appropriate content that demonstrates a positive buying intent.

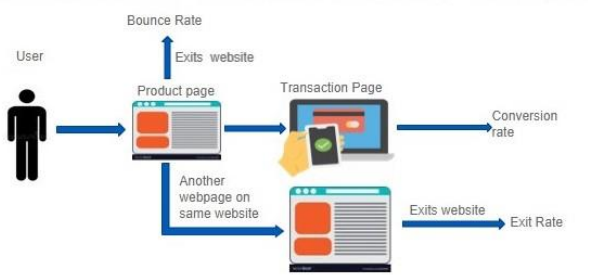


Fig 1. Timeline of a user on e-commerce website

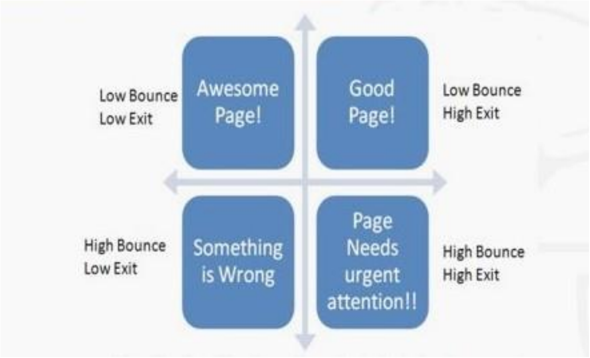


Fig 2. Importance of Bounce Rate and Exit Rate

II. Motivation

In an e-commerce website, user behavior is incredibly significant in determining what content to offer them based on their interests. In this project, we examine client behavior to determine whether they will complete a purchase during that session. The data set consists of the user session data which plays a crucial role in figuring out the user behavior in terms of the time spent on specific page.

Essentially, the idea is to study the user's behavior and then use the clickstream data from that user's web session. Now that we have clickstream data, we can see how many sites the user visited, how much time they spent on each page, and how they arrived at that page.

Now we'll use this clickstream data to figure out why a user isn't completing a purchase transaction, and this will eventually assist the company adjust their website, which will help them raise revenue.

Consumers may easily obtain reviews, ratings, and suggestions before purchasing things due to the massive volume of user-generated content available on digital channels and social media. Consumers could conduct research about products, brands, and enterprises.

Now comes the decision and execution of the purchase on a certain website by the customer. The correct store design, discounts, payment methods, and delivery options go a long way toward preventing customers from abandoning their carts or leaving the website in the middle of their purchase. Our research's data and Ecommerce analytics will eventually assist companies in developing and implementing real-time offers and changes to improve the user's ordering experience.

To address the problem of class imbalance, the training dataset was built from 7000 transactions, half of which were chosen to be from the "purchasing" class. To construct a probability on the consumers' purchasing intention, they used Markov chains, logistic linear regression, decision trees, and Naive Bayes.

To address the problem of class imbalance, the training dataset was built from 7000 transactions, half of which were chosen to be from the "purchasing" class. To construct a probability on the consumers' purchasing intention, they used Markov chains, logistic linear regression, decision trees, and Naive Bayes.

III. INITIAL APPROACH

Initial Approach has been to do the data cleaning followed by feature selection. Feature selection is a process of picking features which add value to the problem statement while performing classification or regression models. Data set consists of 10 numerical variables and 8 categorical variables.

We basically performed Exploratory Data Analysis on our Data set (EDA).

We went on ahead by providing recommendations to improve the website. We trained a variety of classification models namely CART, Logistic Regression, Bagging, Boosting, Random Forest, K-NN and Naïve Bayes and finally we selected the most generalizable model to be Boosting as it had almost same test as well as train Accuracy.

A. Dataset description

The dataset comprises 185,000 Web pages seen by 3500 people over the course of 9800 sessions. The categories of activities ‘‘product view," ‘administrative operation," and ‘‘information acquisition operation" have been derived from the URL information in this dataset. It includes Google Analytics stats for each session, which are useful in determining user involvement. The supplied dataset's essential features are "Page Value," "Exit Rate," and "Bounce Rate”. The time spent on the associated page is also employed as a feature, in addition to the page type information, which is represented by four binary values.

To proceed with the dataset, we must first comprehend the various indicators supplied by Google Analytics.

- Page Value

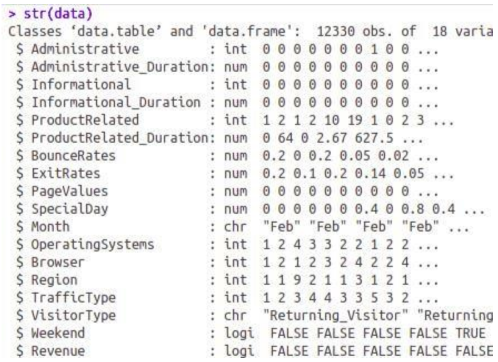
  Page value can be defined as the average number of pages visited by a buyer before finalizing the transaction. This number is meant to help you figure out which page on your site contributed the most to income. Because the page was never visited in a session when a transaction happened, the Page Value for that page will be $0 if it was not engaged in an e-commerce transaction for your website in any manner.

- Bounce Rate

The bounce rate is calculated by dividing the total number of page views by the number of bounces. When a person visits a page and then leaves without viewing another page on the website or engaging with any of the items on the page, this is known as bounce rate.

- Exit Rate

The exit rate is derived by dividing the total number of people who leave your website after seeing a page by the total number of individuals who have seen that page. It is the proportion of visitors who came last in the session, whereas bounce rate is the percentage of visitors who came first in the session.



```
> str(data)
Classes ‘data.table’ and 'data.frame':  12330 obs. of  18 varia
 $ Administrative         : int  0 0 0 0 0 0 0 1 0 0 ...
 $ Administrative_Duration: num  0 0 0 0 0 0 0 0 0 0 ...
 $ Informational          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Informational_Duration : num  0 0 0 0 0 0 0 0 0 0 ...
 $ ProductRelated         : int  1 2 1 2 10 19 1 0 2 3 ...
 $ ProductRelated_Duration: num  0 64 0 2.67 627.5 ...
 $ BounceRates            : num  0.2 0 0.2 0.05 0.02 ...
 $ ExitRates              : num  0.2 0.1 0.2 0.14 0.05 ...
 $ PageValues             : num  0 0 0 0 0 0 0 0 0 0 ...
 $ SpecialDay             : num  0 0 0 0 0 0 0.4 0 0.8 0.4 ...
 $ Month                  : chr  "Feb" "Feb" "Feb" "Feb" ...
 $ OperatingSystems       : int  1 2 4 3 3 2 2 1 2 2 ...
 $ Browser                : int  1 2 1 2 3 2 4 2 2 4 ...
 $ Region                 : int  1 1 9 2 1 1 3 1 2 1 ...
 $ TrafficType            : int  1 2 3 4 4 3 3 5 3 2 ...
 $ VisitorType            : chr  "Returning_Visitor" "Returning
 $ Weekend                : logi  FALSE FALSE FALSE FALSE TRUE
 $ Revenue                : logi  FALSE FALSE FALSE FALSE FALSE
```

Fig. 3. Description of the features

B. *Evaluation metrics*

- Confusion Matrix

A confusion matrix is an important metric used for classification models in machine learning. The predicted values are reflected by the rows and the actual values in the columns. Each element of the confusion matrix tells us how well our model predicts on a given data. We denote these elements by the following names: True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN).

1. True Positive (TP) = The model predicted positive, and it is true.
2. False Positive (FP) = The model predicted positive, but it is not.
3. False Negative (FN) = The model predicted negative, but it is not.
4. True Negative (TN) = The model predicted negative, and it is negative.

- Accuracy

The accuracy of a model tells us how well the model can discriminate between the classes in the testing dataset. The sum of True Positives (TP) and True Negatives (TN) gives us the total predicted correct labels and when divided by the total number of labels(N) gives us the proportion of predicted correct labels.

Accuracy = TP+TN/TP+TN+FP+FN

- Precision

The precision of the model is the fraction of predicted correct labels (TP) to the sum of True Positives (TP) and False Positives (FP).

Precision = TP/TP+FP

- Recall or Sensitivity

The recall is the ratio of predicted correct labels to the sum of True Positives (TP) and False 11 Negatives (FN) or a total number of predicted labels. The recall is also referred to as sensitivity.

Sensitivity = TP/TP +FN

- Specificity

Specificity of a model is the ratio of actual negatives that are correctly identified with the sum of True Negatives (TN) and False Positives (FP).

Specificity = TN/TN+FP

- Feature Selection

Feature selection is basically the process of selecting or choosing variables that are useful for predicting the dependent variable. There are several methods available for the feature selection:

1. Using Correlation between independent and dependent variables.
2. Using Recursive Feature Elimination (RFE) Algorithm
3. Using variable Importance for objects produced by train methods, ‘varImp’ method.

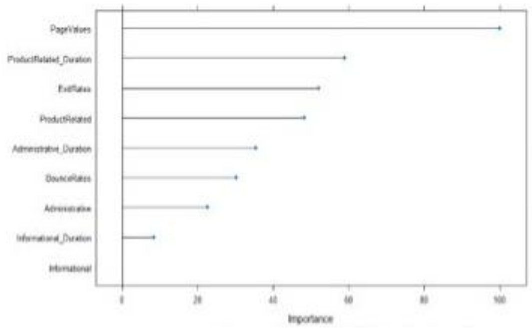
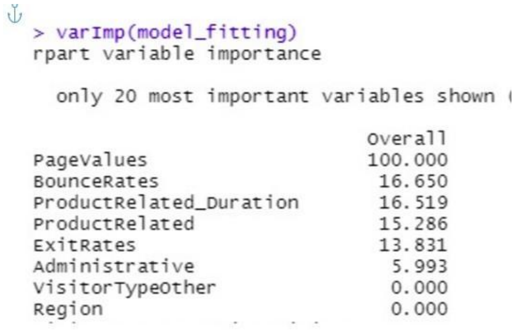


Fig. 4. Features with highest correlation with the Revenue Class

- Variable importance

Variable importance basically refers to how much given model uses that variable to make accurate predictions. Also, more a model relies on a variable to make predictions, more important it is for the model.



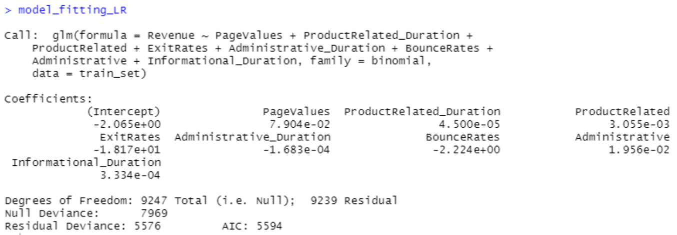
```
> varImp(model_fitting)
rpart variable importance

  only 20 most important variables shown (

                              Overall
PageValues                    100.000
BounceRates                    16.650
ProductRelated_Duration        16.519
ProductRelated                 15.286
ExitRates                      13.831
Administrative                  5.993
VisitorTypeOther                0.000
Region                          0.000
```

Fig. 5. Using varImp for the CART model

IV. ANALYSIS TECHNIQUES

A. *Logistic Regression:*

Logistic regression is a classification technique used to predict a binary dependent variable.

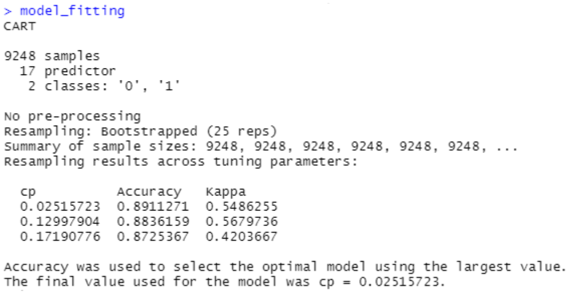


```
> model_fitting_LR

Call:  glm(formula = Revenue ~ PageValues + ProductRelated_Duration + 
    ProductRelated + ExitRates + Administrative_Duration + BounceRates + 
    Administrative + Informational_Duration, family = binomial, 
    data = train_set)

Coefficients:
            (Intercept)               PageValues  ProductRelated_Duration           ProductRelated  
             -2.065e+00                7.904e-02                4.500e-05                3.055e-03  
              ExitRates  Administrative_Duration              BounceRates           Administrative  
             -1.817e+01               -1.683e-04               -2.224e+00                1.956e-02  
 Informational_Duration  
              3.334e-04  

Degrees of Freedom: 9247 Total (i.e. Null);  9239 Residual
Null Deviance:      7969 
Residual Deviance: 5576        AIC: 5594
```

Fig. 6. Logistic Regression model

B. *CART*

CART stands for Classification and Regression Trees. It is more generally known as Decision Trees. It can be used for both regression and classification problems. Here we have used CART for classification.

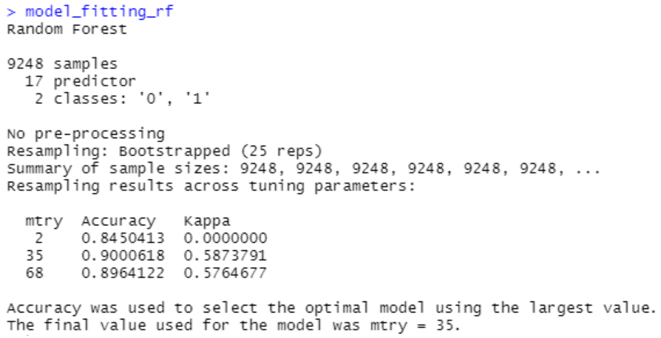


```
> model_fitting
CART 

9248 samples
  17 predictor
   2 classes: '0', '1' 

No pre-processing
Resampling: Bootstrapped (25 reps) 
Summary of sample sizes: 9248, 9248, 9248, 9248, 9248, 9248, ... 
Resampling results across tuning parameters:

  cp          Accuracy   Kappa    
  0.02515723  0.8911271  0.5486255
  0.12997904  0.8836159  0.5679736
  0.17190776  0.8725367  0.4203667

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.02515723.
```

Fig. 7. CART model

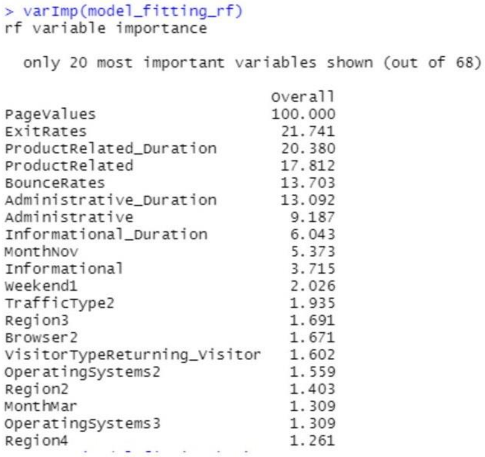
C. *Random Forest*

The purpose of random forest is to overcome the over-fitting problem of individual decision trees by averaging numerous deep decision trees trained on various regions of the same training set.



```
> model_fitting_rf
Random Forest 

9248 samples
  17 predictor
   2 classes: '0', '1' 

No pre-processing
Resampling: Bootstrapped (25 reps) 
Summary of sample sizes: 9248, 9248, 9248, 9248, 9248, 9248, ... 
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa    
   2    0.8450413  0.0000000
  35    0.9000618  0.5873791
  68    0.8964122  0.5764677

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 35.
```

Fig. 7. Random forest model



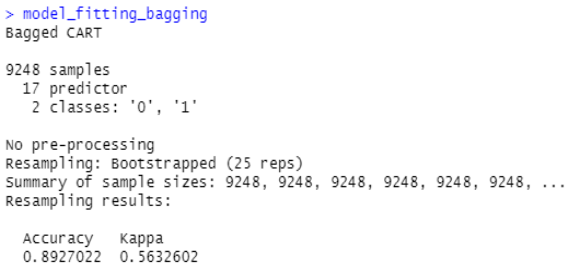
```
> varImp(model_fitting_rf)
rf variable importance

  only 20 most important variables shown (out of 68)

                             Overall
PageValues                   100.000
ExitRates                     21.741
ProductRelated_Duration       20.380
ProductRelated                17.812
BounceRates                   13.703
Administrative_Duration       13.092
Administrative                 9.187
Informational_Duration         6.043
MonthNov                       5.373
Informational                  3.715
Weekend1                       2.026
TrafficType2                   1.935
Region3                        1.691
Browser2                       1.671
VisitorTypeReturning_Visitor   1.602
OperatingSystems2              1.559
Region2                        1.403
MonthMar                       1.309
OperatingSystems3              1.309
Region4                        1.261
```

Fig. 8. Using varImp on Random Forest model

D. *Bagging:*

Bagging is a classification technique that is used to improve the model accuracy by getting an accumulated value from multiple subsets of a dataset. It can be seen as a process in which the original data provided is bootstrapped to make several different datasets or samples and then these multiple samples are taken repeatedly with replacement according to a uniform probability distribution and these models are aggregated by using their average or weighted average.

Using Bagging usually gives us a better balance between potential bias and variance, and it is very useful for nonlinear models.



```
> model_fitting_bagging
Bagged CART

9248 samples
  17 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 9248, 9248, 9248, 9248, 9248, 9248, ...
Resampling results:

  Accuracy   Kappa
  0.8927022  0.5632602
```

Fig 9. Bagging model

V. RESULTS AND OUTCOMES

A. Comparison between models

i. It can be seen from the comparison table given below that we have trained our data set on both Train set as well as Test set. We have also created 3 different tables for Accuracy, Sensitivity and Specificity.

ii. With testing accuracy of 90.34% and training accuracy of 90.10%, Boosting is a highly generalizable model. This is because the model does not overfit and gives low variance.

iii. Also, it has good ability to correctly classify true positives (revenue = Yes) as can be seen from the sensitivity of 96.47%.

TABLE I. Comparison table for different models

	Accuracy		Sensitivity		Specificity	
	Train	Test	Train	Test	Train	Test
Logistic Regression	88.53%	88.54%	97.72%	79.17%	91.38%	93.02%
CART	89.20%	88.81%	95.79%	96.12%	53.18%	48.85%
Bagging	99.82%	90.30%	99.96%	96.24%	99.02%	57.86%
Boosting	90.58%	90.85%	95.91%	96.74%	61.50%	58.70%
Random Forest	100%	90.53%	100%	96.51%	100%	57.86%
K-NN (k=9)	87.96%	86.44%	98.49%	98.20%	30.47%	22.22%
Naive Bayes	84.59%	84.56%	100%	99.96%	0.004%	0.004%

B. ROC Curve for Boosting model

From the receiver operating characteristic curve (ROC) curve shown below it can be seen that Area under the curve (AUC) is reasonable in the sense that it can correctly classify most of the data.

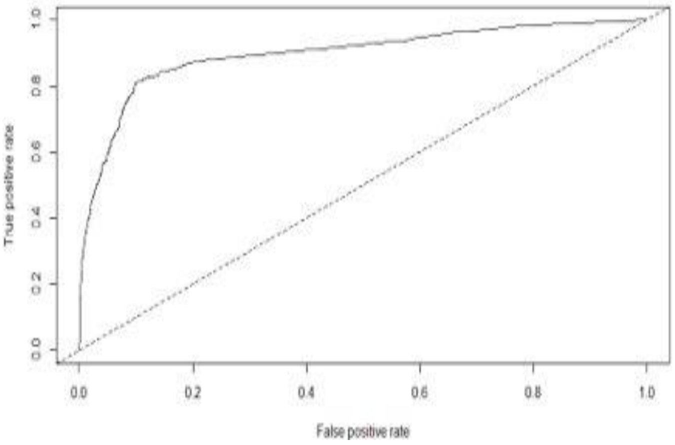


Fig. 10. ROC curve for bagging model

C. Relative influence graph for Boosting

From the figure given below it can be seen that the relative influence of different independent variables on the Revenue class.

Since Page Values has the highest influence, we can recommend that the landing pages can be optimized.

As Exit Rate, Bounce Rate and Product related pages also have high influence we can recommend making an exit strategy with personalized pop-ups and revamping product related pages with a user-friendly UI.

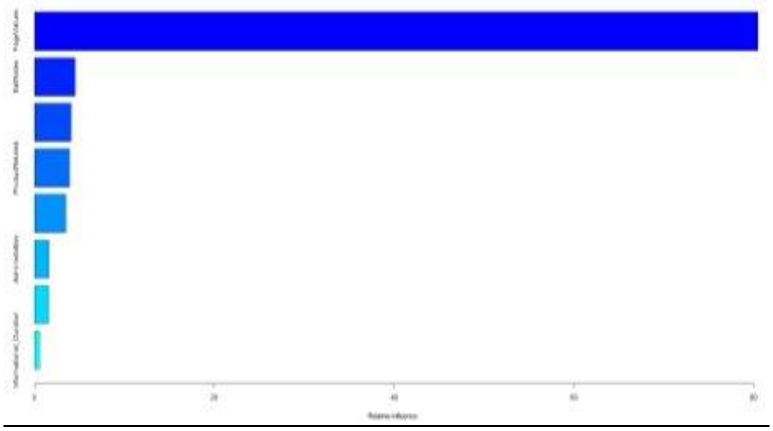


Fig. 11. Relative influence graph

D. Insights obtained after EDA along with recommendations

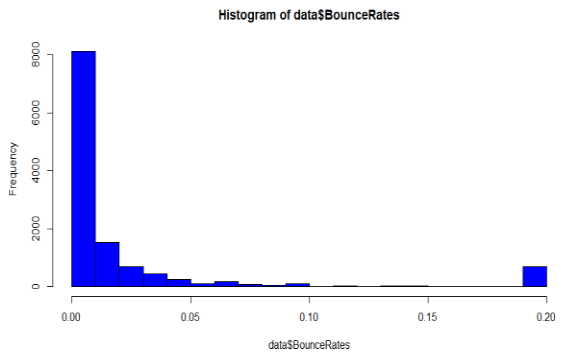


Fig. 12. Histogram of Bounce Rates

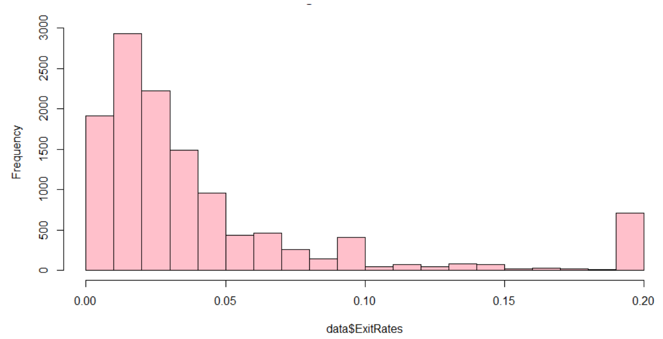


Fig. 13. Histogram of Exit Rates

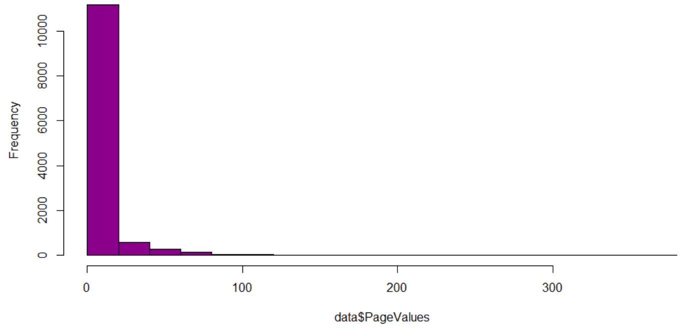


Fig. 13. Histogram of Page Values

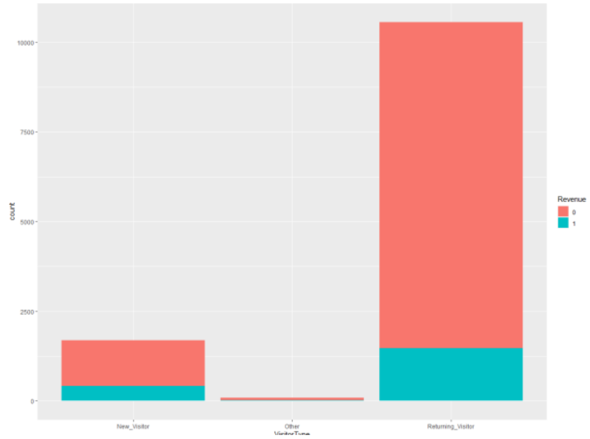


Fig. 14. Visitors type graph

From Fig. 14 we can provide the following recommendations:

- Provide a discount for existing customers to increase conversion.
- Add promotional codes and offers to engage new customers.

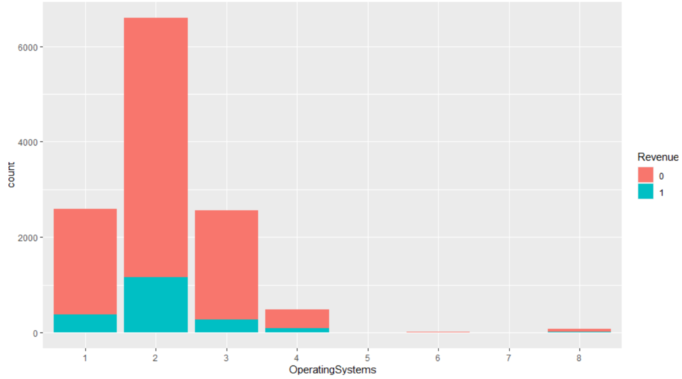


Fig. 15. Graph for type of operating system used

From Fig. 15 we can provide the following recommendations:

- Design the website compatible with all available OS and their versions.
- Add bootstrap and a stylesheet to make the website responsive regardless of the operating system.

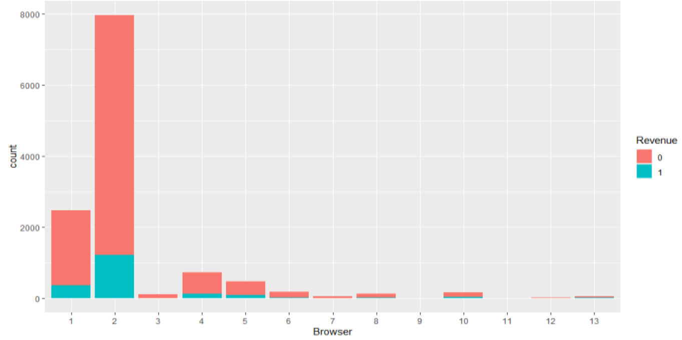


Fig. 16. Graph for type of browser used

From Fig. 16 we can provide the following recommendations:

- Ensure smooth technical operations with a more tailored UI experience that works across all browsers.

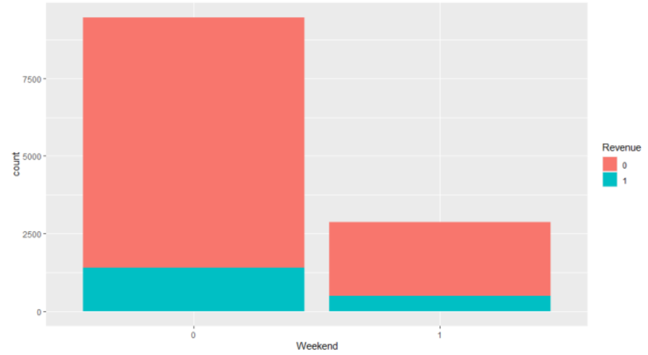


Fig. 17. Graph for weekends and weekdays

From Fig. 17 we can provide the following recommendations:

- Promotional events to engage customers more on the weekends.
- Create discounts specific to weekdays to increase weekday conversion.

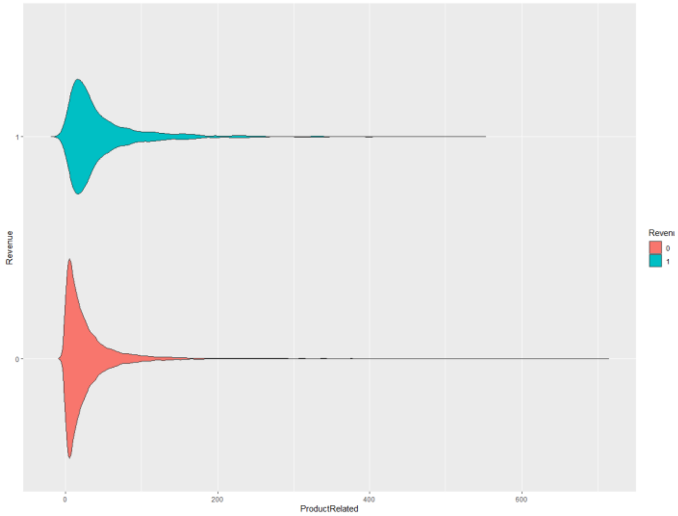


Fig. 18. Graph for type of operating system used

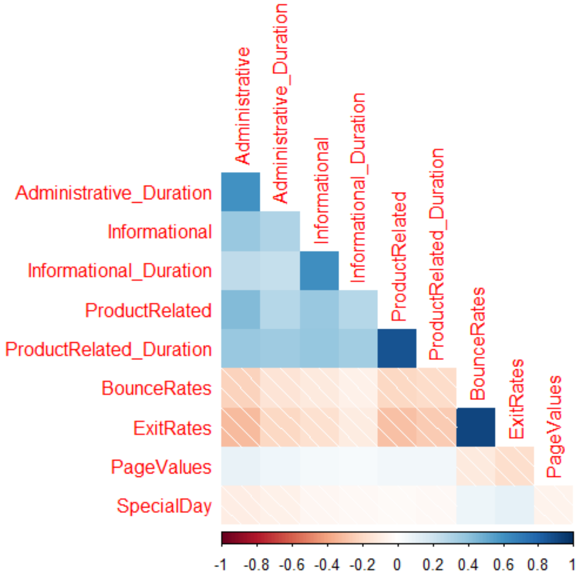


Fig. 19. Correlation Graph

From the correlation graph in Fig. 19 we can give the following inferences:

A. To decrease bounce rate, add pop-ups offering discounts or personalized queries when a customer tries to leave the website.

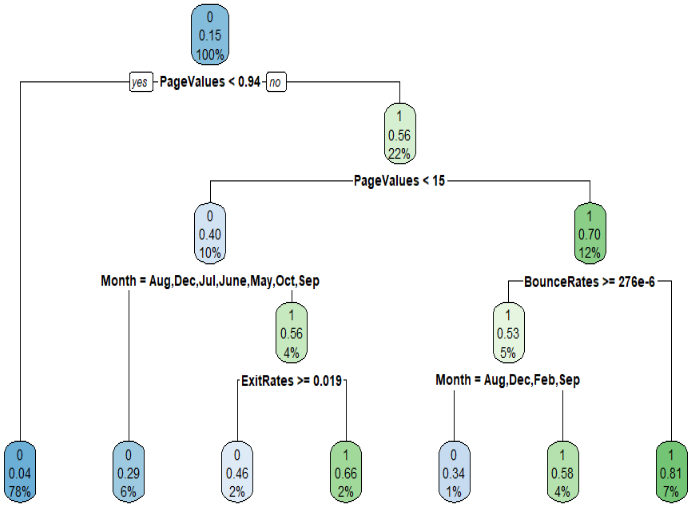


Fig. 20. Decision Tree for our model

From the decision tree graph in Fig. 20 we can give the following inferences:

A. As we see that the page value has the highest impact the landing pages should be optimized to improve the user interaction.

B. The management team should develop a exit rate strategy as the exit rate also plays a important role as can be seen in the graph.

C. The website should have smooth operations in order to lure customers and provide various perks for the newcomers.

D. As we can see the months of August, December, June, July, May, October and September have a much higher purchasing intention than the other months, the company should target this time period to offer discounts and other perks.

VI. BROADER IMPACTS OF OUR RESEARCH

The main motive in our Outcome assists in detecting blind spots to improve conversion rates. Our recommendations enable the companies to make specific adjustments to the website. It also enables the companies to provide consumers with appropriate content that demonstrates a positive buying intent.

In the future, an item or user-based recommender system might be implemented into the system to boost conversion rates even more by providing user-specific material to users who come to the site with the intent to buy and are likely to leave within a prediction horizon. Our findings might be paired with a sentimental analysis of user reviews of a certain website, allowing us to gain a better understanding of the users' purchasing intentions.

Moving forward, we'd like to create an integrated system that combines the project we worked on, which was the prediction of online shopper behavior, with the primary goal of determining a user's purchase intention and assisting businesses in increasing their conversion rate, and sentiment analysis of the company's real-time user reviews, which provides insights into the company's pros and cons. Companies can then concentrate on the primary issues stopping them from prospering.

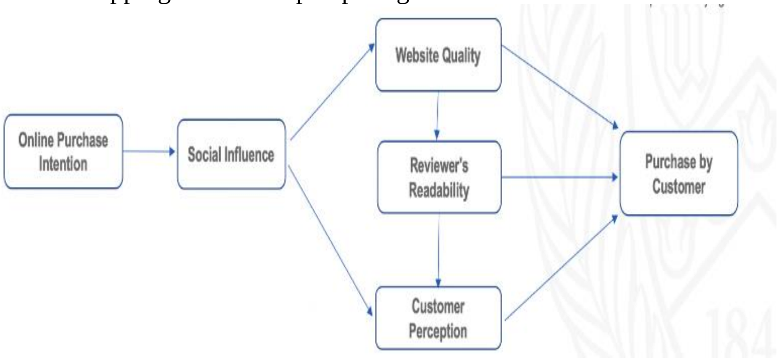


Fig. 21. Impacts of our work

A. *Key takeaways from this project which serve as a base for future research:*

Our research focuses on identifying the blind spots and recommending the companies regarding the issues to be resolved for better performance in terms of revenue and visibility to the users.

Our result is a mix of multiple graphical representations of the user clickstream data of the user's online session, which aided in drawing critical insights into the many parts of a firm such as the influence on the operating system or web browser taken into account while designing for users, the type of emphasis on current users to keep them engaged with the websites, and so on.

B. *Focusing elements for future research*

The future research for this project must involve an in-depth analysis and prediction of a user's accurate purchase intention system based on clickstream data from the user's online sessions as well as comments and evaluations from current and new users.

Data driven model which reuses the historical data which keeps adding on to the existing data and predicts the output based on the accumulated data.

Identifying the major reasons that contribute to a higher bounce rate and exit rate with more precision, such as the user being dissatisfied with the product presentation or website presentation, or the lack of appropriate payment options that are easy for the user.

C. *Outcomes*

Better understanding of the factors influence on the purchasing intent also plays a crucial role. Deeper knowledge of dependency and independent nature of the factors and inter-relation contributes to the research.

Sentiment analysis of comments should also be integrated into the present system to improve its capacity to detect moreaccurate factors influencing the overall path through determining the company’s profit.

D. *Outcomes from sentiment analysis*

1. Contrasting customer feedback with that of competitors
2. Get the most up-to-date product information in real time.
3. 24/7 Hundreds of hours of manual data processing can be saved.
4. Recognize what aspects of your product your clients enjoy and dislike.

SUMMARY

We proposed a real-time online shopper behavior analysis solution in the specified project. As we all know, the rise of e-commerce in recent years has increased market potential, but the fact that conversion rates have not kept up necessitates the development of solutions that provide targeted incentives to online buyers. We used clickstream data from an online store to solve this challenge, which includes numerous Google Analytics metrics (bounce rate, exit rate and page values) for web pages as well as session and user information.

This followed by the verification of the data for missing values, null values, and duplicate values as part of the preprocessing. In addition, we encoded categorical variables by converting them to factors. Then we conducted preliminary exploratory data analysis before beginning to develop the model to learn more about our dataset and how its variables (dependent and independent) are connected or correlated with one another, as well as their overall impact on the dataset.

We developed several graphs based on some of these variables and identified some key insights that can help the organization increase income while maximizing time and resources. We produced several recommendations based on our findings that, if followed, might increase income.

To continue, we used VarImp to do feature selection and choose features that have the most impact on the dataset. We also used correlation plots to determine whether the variables were multicollinear and which were highly associated.

We divided our dataset into a training set and a testing set in an x:y ratio and used the training set to train multiple machine learning models. Logistic regression, CART (Classification and Regression Trees), Bagging, Boosting, Random Forest, Naive bayes and KNN (K-Nearest Neighbors) were all employed.

We gathered key model assessment metrics such as (Accuracy, Sensitivity, and Specificity) after testing these models on test and train data and offered a comparison table of these metrics for all the algorithms listed.

The boosting model produces the best results since it has a little difference in accuracy between training and testing data and its sensitivity is extremely accurate. The ROC Curve for the boosting model was plotted to demonstrate why boosting was chosen as our generalizable model.

REFERENCES

[1] Online Shoppers Intention UCI Machine Learning Kaggle

[2] Online Shoppers Purchasing Intention Dataset Data Set University of California Irvine

[3] Sakar, C.O., Polat, S.O., Katircioglu, M. and Kastro, Y., 2019. Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications*, *31*(10), pp.6893-6908.

[4] Ding, A.W., Li, S. and Chatterjee, P., 2015. Learning user real-time intent for optimal dynamic web page transformation. *Information Systems Research*, *26*(2), pp.339-359.

[5] Mobasher, B., Dai, H., Luo, T. and Nakagawa, M., 2002. Discovery and evaluation of aggregate usage profiles for web personalization. Data mining and knowledge discovery, 6(1), pp.61-82

[6] Fernandes, R.F. and Teixeira, C.M., 2015. Using clickstream data to analyze online purchase intentions.

[7] Suchacka, G., Skolimowska-Kulig, M. and Potempa, A., 2015. Classification Of E-Customer Sessions Based On Support Vector Machine