Online Shoppers' Purchasing Intention

Anuj Singhvi *Dalarna University v22anujs@du.se*
Pasumarthy Krishna Bharat *Dalarna University h21pasbh@du.se*

Abstract – In this paper, we analyze the "Online Shoppers Purchasing Intention Dataset" to predict whether a user session will culminate in a successful purchase. We investigate which type of user sessions lead to a purchase and examine variables that contribute to the revenue field. We further provide a set of recommendations that the companies can consider to increase their sales. The dataset consists of a certain number of session-related and user-related information. For exploratory data analysis, data visualization, and predictive analysis we have used Python and its libraries. For preparing the dashboard we have used Power BI Desktop and published the dashboard to Power BI Service. In our study, we made the predictive analysis through K Nearest Neighbors, Random Forest Classification, Support Vector Machine Classification, and Logistic Regression. We achieved prediction with 88% accuracy in Random Forest Classifier model that whether a user will make a final purchase or not. We concluded that the shorter the sessions, the more likelihood of making a final purchase by the user. We analyzed the Revenue field through different variables like 95% of the successful sales are made during Special Days. Basis the insights, we also made recommendations towards making targeted marketing and promotional activities like discounts, deal of the day offers, etc. to the users that can aid companies to grow their sales.

Keywords: Online Shopping, E-commerce, Purchase Intentions, Predictive Analysis, KNN, Random Forest, Support Vector Machines, Logistic Regression, Revenue

I. INTRODUCTION & LITERATURE REVIEW

The number of people who buys online (digital buyers) is only expected to go up in future and is currently estimated to be over 2 billion in 2021 [1]. Online purchases also got accelerated recently because of the social distancing and health norms imposed due to the COVID pandemic. A latest McKinsey report states that in the US, the pandemic has changed the behaviors of shoppers and more consumers are anticipated to make a proportion of their purchase online even after COVID [2]. Therefore, in today's world of e-commerce and online shopping, it is critical for businesses to know about the intention of users towards buying their products on their websites.

The growth in online buyers does not necessarily translate to these buyers making final purchases [3-5]. One of the statistics shows that the conversion rate of a visit to a successful online purchase was a mere 2% for online shoppers worldwide for the third quarter of the year 2020 [6]. Another estimate from SmartInsights echo a similar story of out of the total online sessions, only 43% move to the product page view and a meager 3.3% goes towards a final transaction [7].

There can be many reasons why customers are dropping off. One of them is that they do not have any buying intentions and may be there just to browse the catalogues. The second one can be that the landing or product page has issues with poorly designed User Interface (UI) or User Experience (UX) that consumes a lot of time to load or may not have a clear Call To Action (CTA) [8]. Moreover, users can also abandon the products during the final checkout stage due to high delivery/transaction costs involved or if they find the payments failing after multiple attempts [9].

It is also a costly affair to lose existing customers. One statistic points towards the fact that the cost of acquiring a new customer is as much as five times to retain an existing one [10]. From the above discussions, it is clear that businesses have challenges both with respect to retaining existing customers and acquiring new ones. Consequently, they should have consistent efforts to keep converting their website traffic to final purchases.

There are some studies conducted that focusses on how to improve the purchase conversion rates and shopping cart abandonment [11-13]. The objective of this project is to determine whether a user session on a website will culminate in a final purchase or not. To execute on this objective, we have generated the below three research questions (RQ):
RQ1. Which kind of sessions, viz. Administrative, Informational, Product related lead to a purchase?
RQ2. Determine whether a user will make a purchase or not through predictive analysis
RQ3. Through visualization, identify insights on the successful purchases based on the types of visitors, purchase months, the occurrence of special days

The rest of the report is divided into the following sections – Section II spells out the Materials and Methods deployed to answer the research questions. Section III presents the results and Section IV provides the discussions of the results. Section V finally mentions the conclusion of this paper along with the limitations of the work.

II. MATERIAL & METHODS

The dataset is titled "Online Shoppers Purchasing Intention Dataset" and is available at UCI. It comprises feature vectors that relate to 12,330 sessions and each of these sessions is of a unique user for a one-year period [14]. There are 18 feature vectors with 10 of them being numerical and 8 categorical. The "Revenue" field shows whether a successful purchase was made or not. Out of 8 categorical variables, 4 variables are already

converted to numeric values and information regarding these 4 variables is not disclosed in the description. For example, a feature named Region has values from 1 to 9 in the dataset, but there is no information regarding what each number is referred to. An excerpt of the dataset is provided below and detailed information on the variables is provided in the Appendix (Table 1).

Administrative	0
Administrative_Duration	0.0
Informational	0
Informational_Duration	0.0
ProductRelated	1
ProductRelated_Duration	0.0
BounceRates	0.2
ExitRates	0.2
PageValues	0.0
SpecialDay	0.0
Month	Feb
OperatingSystems	1
Browser	1
Region	1
TrafficType	1
VisitorType	Returning_Visitor
Weekend	False
Revenue	False

Figure 1: Dataset Excerpt

Exploratory Data Analysis:

To execute on the first research question on analyzing the diverse kinds of sessions that will lead to purchase, we shall undertake the below steps. After we import the key libraries and the CSV file, we explore the dataset using *describe* and *info* functions. To see how the variables are correlated with each other, we run the correlation function. We observe some expected correlation between Administrative session and its duration and similarly for Information and ProductRelated fields. There is one interesting relation that we want to highlight - between Revenue and PageValues at ~0.5 (Figure 2)

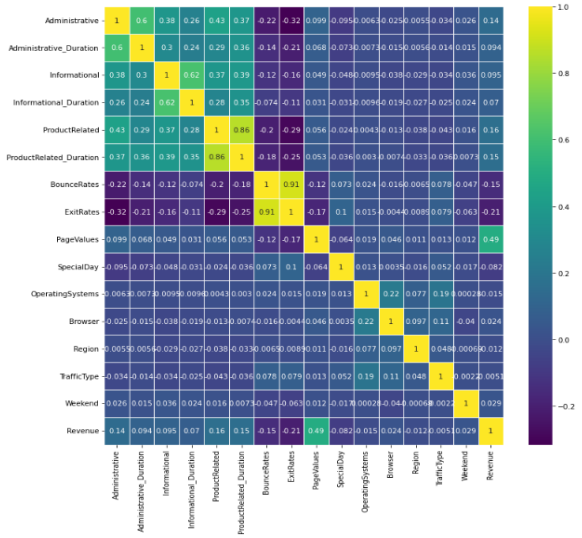

Figure 2: Correlation table of the variables

Data Visualization:

We shall now explore our important observations from the research. Since we had a significant correlation between Revenue and Page-Values, we shall analyze the same through the lens of the Revenue field. Simply put, the "Page Value" tells us about the average value for a web page that a user visited before completing the transaction

Data Preprocessing for Modelling:

Through the *.isna* function, we check that there are no NA values that we are dealing with. Further, we check on the duplicates present in the dataset which comes out to be 125. We delete the duplicates using *drop_duplicates* function.

We clean or manipulate the data primarily to convert the required columns to numeric. The variables that are converted are listed below:

- Administrative Duration, Informational Duration, Product Related Duration, Bounce Rates, Exit Rates, Page Values columns are float type with higher decimal points. They are restricted to 1 (one) decimal place.
- Categorical columns Month, Operating Systems, Browser, Region, Traffic Type, Visitor Type, Weekend columns are encoded using the One Hot Encoding package and transformed the data from 18 columns to 76 columns.
- Lastly, converted the data type Revenue to a category as it is the response variable.

After dropping the duplicates and the above transformations, we have 12205 rows out of which we have 10297 rows as 0 (False - Not Purchased) and 1908 rows as 1 (True - Purchased) for the response variable. The count of rows with revenue as False is over 5 times than the count of rows with response as True.

Data Balancing:

To avoid inaccurate results due to imbalances in the dataset, under sampling was performed to have a ratio of about 3:1. Hence, after randomly dropping 4000 rows with 'Revenue' field as 'False', the dataset contains 6297 rows with 'False' as response and 1908 rows with 'True' as the response. Equally balanced dataset was not preferred as the possibility of having a non-purchase transaction is more. Consequently, having that little imbalance in the dataset will let the model consider the imbalance as well.

Predictive Analysis:

To perform the predictive analysis for the classification problem to identify if a transaction ended up with or without a purchase, the following 4 machine learning techniques were implemented.

1. K Nearest Neighbors
2. Random Forest Classification
3. Support Vector Machine Classification
4. Logistic Regression

Data which is created after the preprocessing is used directly for KNN and Random Forest models. Scaled data applied on Principle Component Analysis with 0.95 percent of variance ratio which resulted in 52 Principal Components which are used for SVM and Logistic regression models. Let us delve deeper into the models now.

1. **K Nearest Neighbor:**

Model Highlight (used for evaluation): Unscaled dataset - 80% Train & 20% Test - K = 3 Neighbors - 5-Fold Cross Validation - KNN Model

For identifying the significant value of K, we performed iterative modeling with different values of K from 1 to 10 and plot the accuracy results against value of K. We have decided to proceed further with K = 3 as the accuracy of test data is almost close to the highest accuracy in as per the plot. Even the train dataset has a better accuracy for K = 3 compared to 5,7 and so on. We performed K-Fold cross validation with K as 5 to train the model.

2. **Random Forest Classification:**

Model Highlight: Unscaled dataset - 80% Train & 20% Test - 1000 Iterations - 5-Fold Cross Validation - RFC Model

To execute, we considered 1000 iterations by setting 'n_estimators' parameter to 1000 for the Random Forest classifier model. We then ran the K-Fold cross validation with K as 5 to train the model.

3. **Logistic Regression**

Model Highlight: Scaled dataset – PCA (95%) - 80% Train & 20% Test - 5-Fold Cross Validation - Logistic Regression Model

We executed K-Fold Cross Validation with K as 5 to train the Logistic regression model

4. **Support Vector Machine**

Model Highlights: Scaled dataset – PCA (95%) - 80% Train & 20% Test - RBF Kernel - 5-Fold Cross Validation - Support Vector Machine Model.

We created a Support Vector Classifier using a Radial Basis Function (RBF) kernel. We now test with K-Fold cross validation with K as 5 to train the Support Vector Machine model.

Data Analysis:

For the research problem 3 for creating the live dashboard, we have used Power BI Desktop. Using the Power Query Editor, we have transformed the data to convert the numeric columns to categorical Drop unnecessary columns and created dashboards for analyzing the behaviors of features like Visitor Type, Region, Month, Special Days, Weekends with respect to Revenue

III. RESULTS

RQ1: Analyze different sessions behavior with respect to revenue.

We plot (Figure 3) the information for each of the session duration and observe a visibly similar trend. The shorter the duration, the more the purchases conversion and, conversely, the more the interaction time increases the lesser the final conversion.

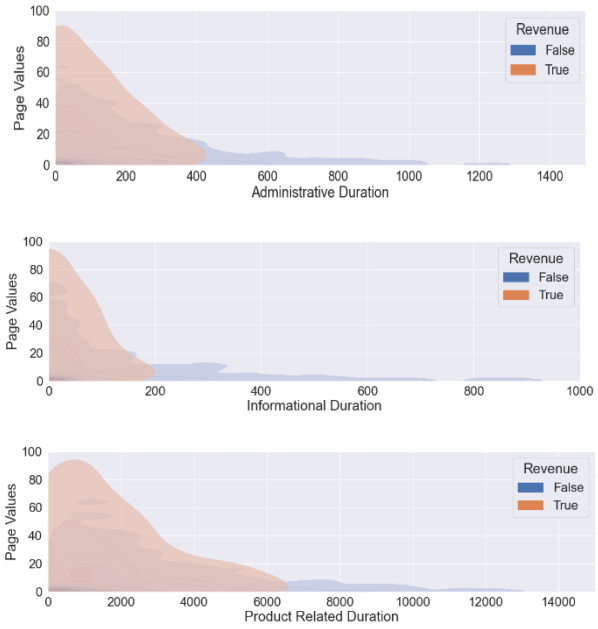

Figure 3: Revenue distributed over session durations and PageValues

RQ2: Perform predictive analysis to know if the purchase was made or not.

Based on the validation set approach, out of 8205 rows obtained after under sampling, we have 6564 rows in train dataset and 1641 rows in the test dataset. The dataset was split randomly using the test train split function in Sklearn Model Selection Package.

K Nearest Neighbours:

The KNN model with K as 3 was performing well for False Negative values with 1135 being predicted correctly and only 121 were predicted wrong. But for True positive values only 167 were classified correctly and 218 are predicted as false. We have also observed that with increasing values of K accuracy was decreasing.

Random Forest Classifier:

The Random Forest Classifier model with 1000 iterations has a better performance compared to KNN for False Negative values with 1177 being predicted correctly and only 79 were predicted wrong. For True positive values 265 were classified correct and 120 are predicted as false which is far better compared to KNN.

Logistic Regression:

The Logistic Regression Model has a better performance compared to KNN and Random Forest for False Negative values with 1218 being predicted correctly and only 38 were predicted wrong. But for True positive values 183 were classified correctly and 202 are predicted as false which is far better compared to KNN but not as much as Random Forest Classifier.

Support Vector Machine:

The Support Vector Classifier model with radial basis function

kernel has a better performance compared to KNN and Random Forest for False Negative values with 1215 being predicted correctly which is almost close but less than Logistic Regression and only 41 were predicted wrong. For True positive values 197 were classified correct and 188 are predicted as false which is better compared to KNN and Logistic but lesser than Random Forest.

Below tables show the comparison of the accuracy, precision and recall scores of all 4 models for train and test data.

Train Data Scores (in %)			
Model	Accuracy	Precision	Recall
K Nearest Neighbours	88.15	80.86	64.08
Random Forest Classifier	100	100	100
Logistic Regression	84.99	79.89	47.21
Support Vector Machine	86.52	83.3	52.4

Table 2: Training Data Scores

Test Data Scores (in %)			
Model	Accuracy	Precision	Recall
K Nearest Neighbours	79.34	57.99	43.38
Random Forest Classifier	87.87	77.03	68.83
Logistic Regression	85.37	82.81	47.53
Support Vector Machine	85.5	82.1	48.83

Table 3: Test Data Scores

RQ3: Through visualization, identify insights on the successful purchases

The dashboard that can be utilized by decision makers to visualize and analyze data to aid in decision making. We have built it on Power BI Desktop and then published in Power BI Service (link to view [Dashboard - Power BI)]

Here, the first bar chart (Figure 4) focusses on the Revenues distribution over month from highest to lowest. November and May are the top while June and February feature in the bottom two.

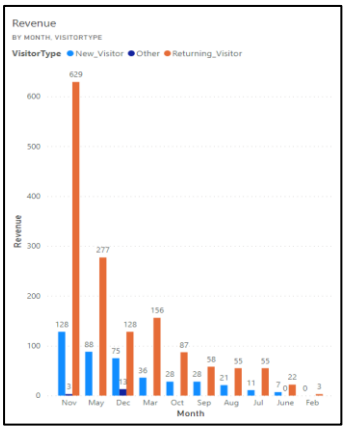

Figure 4: Revenue distribution by months

The next highlight the distribution of Revenue from Visitor Type (Figure 5). The pie chart below is quite telling the fact that 3 out of 4 final purchases come from returning user and only 1 out of 4 is from a new user. To target the new users, the management can offer first-time discounts or free deliveries, etc. Further, they can also make the new users journey smooth and effortless so that the fresh visitors directly go for the final purchase.

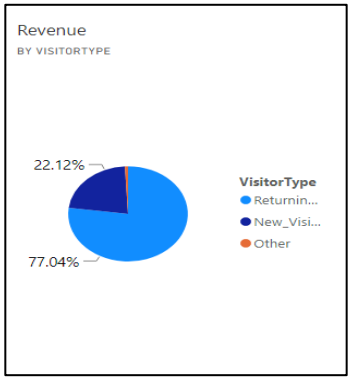

Figure 5: Revenue distribution by VisitorType

We now look at a very striking statistic from Figure 6(a) displaying how are the successful purchases divided between Special Days (like Valentine Day, Mother's Day, etc.). As per the description of the value, 0 is given values to date one day before and after the special day. For example, February 13 and February 15 will have a value of 0 for the special day of February 14th (Valentine's Day). The value comes out to be ~95%. This clearly communicates the fact that almost every successful purchase made on the ecommerce platform is during the special days. It shows that most visitors are crystallizing their purchases during these days.

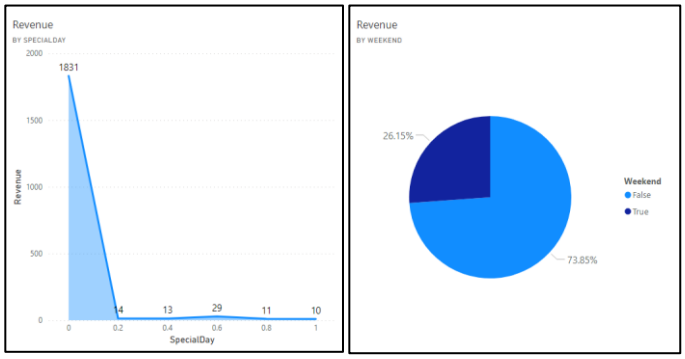

Figure 6 (a & b): Revenue distribution by Special days (a) and Weekend (b)

The final visualization is towards the understanding of breakup of sales made during weekend and non-weekend days (Figure 6(b). Through the pie chart, we can infer that almost 1 out of 4 sales were made during the weekends. This information can be extremely helpful to the management as promotional or marketing campaigns to nudge visitors to buy during weekdays can be implemented. Also, to boost sales during weekends (where visitors may have more time to spend), targeted campaigns can be undertaken.

IV. DISCUSSON

Based on the results of RQ1, one likely reason for quick sessions leading to purchases can be that the users already have a high intention to buy a product and so immediately make the purchase. This can be seen in Figure 3 where the part on the left-hand side (red area) is for purchases and the right one (blue) is for non-purchases.

Based on the results of RQ2 (Table 2 and Table 3), for the Train dataset Random Forest Classifier has 100 percent accuracy, precision and recall where over fitting of the model can be a reason. KNN has better precision and recall percentage compared to the other 2 for the training datasets. But for Test dataset Random Forest has the highest accuracy followed by SVM and Logistic regression. Logistic regression has the best precision followed by SVM and Random Forest Classifier. Random Forest has the highest recall score compared to all others which had less than 50%.

For the insights gained in RQ3 through visualizations, we recommend the following actions to companies' management:

- To pump up monthly sales, offers like month-end deals or pay-day/salary-day sales can be initiated. This can prompt the users to not delay their purchases to special days / months and finalize during last calendar month days or salary day (like, 25th of every month in Sweden).
- To acquire new users and increase their share in final purchases, first-time discounts, waving delivery/service fees, and free 30-day return styled promotion campaigns can be targeted.
- For non-special / non – occasion days, companies need to make their website more attractive to users. It can be done through 'deal of the day' offers or running brand specific promotions. Another way to activate the users during this time is to remind them of their items in their Wishlist or final Shopping Carts.
- During weekdays, users may have a limited time to purchase. So, on the spot or lightning discounts can be notified to customers in real-time to increase the likelihood. For sales during weekends, shoppers have more time and so a live chat pop-up or a quick customer-company call can be set up to seal the purchases.

V. CONCLUSION

As businesses expect to have the online medium as a key driver of their sales, it is imperative for them to understand the buying purchase intentions of the visitors. Through this work, we have made predictive analysis on the dataset that can be deployed by other organizations to estimate whether a user session will convert to purchase or not. We have also analyzed the Revenue field from the lens of different variables of months, visitor types, and presence of special days and weekends to derive insights on the type of session leading to a purchase and provided recommendations that can be deployed by companies to increase sales.

Although we have done predictive analytics with 4 methods, we do acknowledge that we could have also considered other classification techniques including Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Decision Trees, etc. Moreover, we have deployed under sampling in the work. It would be interesting to see the analysis that we could have gotten from balanced and oversampling.

References:

[1] Statista, "Number of digital buyers worldwide from 2014 to 2021," 2022. [Online]. Available: https://www.statista.com/statistics/251666/number-of-digital-buyers-worldwide/

[2] B. C. K. R. a. J. W. Tamara Charm, "The great consumer shift: Ten charts that show how US shopping behavior is changing," 2020. [Online]. Available: https://www.mckinsey.com/business-functions/growth-marketing-and-sales/our-insights/the-great-consumer-shift-ten-charts-that-show-how-us-shopping-behavior-is-changing

[3] Cairnona CJ, Rarnirez-Gallego S, Torres F, Bernal E, del Jes6s MJ, Garcia S (2012) Web usage mining to improve the design of an e-commerce website: OroliveSur. com. Expert Syst Appl 39(12):11243-11249

[4] Rajamma RK, Paswan AK, Hossain MM (2009) Why do shop- pers abandon shopping car t? Perceived waiting time, risk, and transaction inconvenience. J Prod Brand Manag IS(3): ISS-197

[5] Ding AW, Li S, Chatterjee P (2015) Learning user real-time intent for optimal dynamic web page transformation. Inf Syst Res 26(2):339-359

[6] Statista, "Conversion rate of online shoppers worldwide as of 3rd quarter 2020," 2021. [Online]. Available: https://www.statista.com/statistics/439576/online-shopper-conversion-rate-worldwide/.

[7] D. Chaffey, "E-commerce conversion rates benchmarks 2022 – how do yours compare?," 2022. [Online]. Available: https://www.smartinsights.com/ecommerce/ecommerce-analytics/ecommerce-conversion-rates/.

[8] J. BULLAS, "Why Your eCommerce Sales Funnel Isn't Working (And How to Fix It)," 2020. [Online]. Available: https://www.jeffbullas.com/ecommerce-sales-funnel/.

[9] Prefix, "15 Common Online Shopping Problems Causing Revenue Loss for Your Business (+ How To Fix or Avoid Them)," 2017. [Online]. Available: https://www.prefixbox.com/blog/online-shopping-problems/#8b122e8f79e5.
[10] W. Tidey, "Acquisition vs Retention: The Importance of Customer Lifetime Value," 2018. [Online]. Available: https://www.huify.com/blog/acquisition-vs-retention-customer-lifetime-value.
[11] Awad MA, Kha1i1 I (2012) Prediction of user's web-browsing behavior: application of markov model. IEEE Trans Syst Man Cybern B Cybern 42(4):1131-1142
[12] Budnikas G (2015) Computerised recommendations on c-trans- action finalisation by means of machine learning. Stat Transit New Ser 16(2):309-322
[13] Fernandes RF, Teixeira CM (2015) Using clickstream data to analyze online purchase intentions. Master's thesis, University of Porto
[14] Dataset link - archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

Appendix (Table 1)

COLUMN VARIABLES	DESCRIPTION
Administrative	Number of administrative pages like account management visited by user
Administrative_Duration	Time spent (in seconds) on administrative page category by user
Informational	Number of Informational pages like communication and address information of the shopping site visited by user
Informational_Duration	Time spent on Informational page category by user
ProductRelated	Number of product related pages visited by the user
ProductRelated_Duration	Time spent on product related page category by user
BounceRates	Percentage of visitors who enter from that category of page and then leaves without triggering any other requests
ExitRates	The value of "Exit Rate" feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the las in the session.
PageValues	The Page Value feature in Google Analytics represents the average value for a web page that a user visited before completing an e-commerce transaction. In the dataset, "PageValues" column shows the average page values of the pages visited by the user
SpecialDay	It shows how close the browsing day was from special day (e.g., Mother's Day)