Literature Review

Unveiling Patterns in E-commerce: A Predictive Analytics Exploration of Online Shoppers

Md Khaled Saifullah

Student Number: 501230688

Supervisor: Tamer Abdou, PhD

Date of Submission: 27th February 2024

Institution: Toronto Metropolitan University

Table of Contents

Abstract (Revised)

Title:

Unveiling Patterns in E-commerce: A Predictive Analytics Exploration of Online Shoppers Purchasing Intention Data

In the rapidly evolving digital landscape, understanding user behavior within e-commerce platforms is crucial for businesses aiming to optimize strategies and enhance revenue. This study embarks on an exploration of purchasing intention data within the e-commerce domain, employing Predictive Analytics as a guiding methodology. The primary aim is to decode intricate patterns embedded within user engagement data, anticipate user actions, and thereby bolster online business revenue.

The project confronts the formidable task of deciphering complex patterns enshrouded within copious volumes of clickstream data. Five cardinal research questions have been articulated to address this challenge, focusing on aspects such as the influence of different page categories on user purchases and variations in user behavior during weekends and special days. The endeavor is to glean insights into engagement metrics' predictability for transactions and identify distinct user segments predicated on categorical attributes. The project aims to answer the following research questions:

How do different informative page categories (like Informational and Product Related pages) contribute to the likelihood of a user making a purchase?

Can we predict the likelihood of a user making a purchase based on metrics such as Bounce Rates, Exit Rates, and Page Values?

How does user behavior vary on weekends compared to weekdays in terms of engagement and conversion rates?

What is the impact of special days on user engagement and transaction rates?

Since question 3 and 4 both deals with the time of purchase and generating revenue, it might be difficult to distinguish the duration since out dataset is only containing 10 months dataset and weekend-weekdays might overlap with the way the data represent the special events. Rather considering the trend of timings for all together and answering the question as a combined will be more practical and informative for the findings.

Question 3: What is the relationship between features related to timing (Weekend, Month, and special Day) and revenue generation?

This research project will leverage the Online Shoppers Purchasing Intention Dataset sourced from the UC Irvine Machine Learning Repository as the primary dataset for investigative analysis and to address the formulated research questions. This robust dataset, comprising 12,330 instances enriched with 10 numerical and 8 categorical attributes, forms the foundation for this analytical expedition. Attributes encapsulate diverse elements including pages visited, visit durations, bounce rates, exit rates, page values, special day indicators alongside user characteristics like operating systems browsers and regions.

A hybrid analytical approach is adopted to navigate through these research questions, amalgamating exploratory data analysis with statistical tests while leveraging classification models the models will include Decision trees, Random Forests, K-nearest neighbours, and clustering algorithms for deeper insights. Python emerges as the instrumental programming language supported by libraries like Scikit-learn for machine learning applications; Pandas for adept data manipulation; Matplotlib facilitating intuitive visualizations-all encapsulated within Jupyter Notebooks environment ensuring an interactive exploration.

Introduction

The culmination of this study promises actionable intelligence enabling businesses to refine strategies meticulously; elevate user experiences aligning with individual preferences; optimize revenue streams adapting to dynamic market trends-all underpinned by a comprehensive understanding gleaned from e-commerce data analytics. This integration of Predictive Analytics underscores a paradigm shift towards informed decision-making processes illuminating pathways for enhanced business performance in an increasingly competitive digital marketplace.

Research Questions

How do different informative page categories contribute to the likelihood of a user making a purchase?

In the ever-evolving landscape of e-commerce, understanding the factors that contribute to user purchasing decisions is paramount for businesses seeking to optimize their online platforms and enhance the overall user experience. One crucial aspect of this understanding involves exploring the impact of different informative page categories on the likelihood of a user making a purchase.

Different informative page categories, such as informational and product-related pages, can contribute to the likelihood of a user making a purchase. Research has shown that the expression of advertised information can effectively enhance consumers' purchase intention, depending on the product type. Understanding the impact of different informative page categories and tailoring the expression of advertised information accordingly can help improve users' likelihood of making a purchase.

The research by Peter et al. investigates the influence of **product description** on **purchase intention** within the context of online stores on marketplace platforms. The study aims to determine whether product description affects purchase intention. The researchers employ **structural equation modeling** (SEM-PLS) using SmartPLS 3.3.7 software. The research findings reveal that product description has a **positive and significant effect** on both enduring and situational involvement. The results diverge from previous studies that suggested both enduring and situational involvement play a role in influencing purchase intention.

Understanding how different informative page categories -such as product descriptions, impact user behavior is essential for designing effective e-commerce strategies. By recognizing the nuanced interplay between involvement, content, and purchase intent, businesses can tailor their approaches to meet user needs and drive successful conversions.

Wang et al. (2022) in this paper “Promote or inhibit? Research on the transition of consumer potential purchase intention. Annals of Operations Research” used a continuous-time hidden Markov model (CT-HMM) to capture the transfer path of consumers who are affected by various online channels and found that online page view (advertising) has a positive and statistically significant impact on the transition of consumer purchase intention.[2]

Sunarto et al. (2023) [1] paper aims to reduce classification error in online shoppers' purchasing intention using the boosting technique with the C4.5 algorithm. The experimental results show that applying the boosting technique improves accuracy and lowers classification error compared to using the C4.5 algorithm without boosting. The paper develops a real-time system for predicting online shoppers' purchasing intentions by utilizing the Random Forest algorithm and considering both known and anonymous sessions. The results show that the Random Forest algorithm outperforms other methods in terms of accuracy and F1 Score, highlighting its effectiveness in forecasting purchasing intent in an e-commerce environment.

The study conducted by Frazier et al. (2022) [4] focuses on the data analysis of online shopper's purchasing intention, employing machine learning for predictive analytics. The research adopts a comprehensive data analytics workflow, encompassing data ingestion, descriptive statistics, and exploratory data analysis. The authors employ various methods, including frequency, probability histograms, and scatterplots, to examine the impact of numerical features on the target class. The analysis extends to categorical features, exploring their influence on the class label through frequency and probability assessments. We will use the statistical approaches although the data processed prior to this study are different as they use numerical instead of categorical variable.

Kabir et al. (2019) [5] in the paper analyzed different classification algorithms such as Decision Tree, Random Forest, Naive Bayes, and SVM to predict whether a customer will make a purchase or not when visiting the webpages of an online shop. Ensemble methods were also performed to boost the performance of these algorithms. We are planning to use the classification algorithms to find out the different informative page categories impact in purchasing decision. The study analyzed empirical data of online shoppers to build a prediction model for their purchase intention. The application of ensemble methods to the dataset made this study unique.

Kushwaha, Rahul., P., Anudeep., Desai, Shubham, Uday. (2023) “Variable Aware Analytic Driven Online Shoppers Purchasing Intention using ML Algorithms” [3] in their final model achieved a higher accuracy on the test set, indicating that it can accurately classify website visitors with a high degree of accuracy. The Random Forest algorithm outperformed other methods in terms of accuracy and F1 Score, demonstrating its effectiveness in predicting online shoppers' purchasing intentions in an e-commerce environment. The analysis identified the most crucial features for predicting purchasing intentions, including PageValues, Exit Rates, Administrative Duration, ProductRelated Duration, and Bounce Rates. These findings offer valuable insights into the factors that impact the purchasing decisions of website visitors. The study highlights the success of machine learning techniques in anticipating the intentions of internet shoppers and emphasizes the significance of properly preparing data and developing appropriate features to construct precise models. The research also emphasizes the importance of evaluating ML algorithms' performance through different metrics to identify their strengths and weaknesses.

Can we predict the likelihood of a user making a purchase based on visitor type (new, returning, other)? / Is there an impact of visitor type (new, returning, other) on whether revenue will be generated?

The likelihood of a user making a purchase can be predicted based on metrics such as new, returning, other. Descriptive statistical analysis of online shoppers' purchasing intention data showed that variables like time spent and page values are positively correlated with shopping intention, while variables including bounce rate and exit rate are negatively correlated with shopping intention.

In "The Application of Machine Learning in Online Purchasing Intention Prediction" by Xiang, Shi. (2021) [7], the author employs a machine learning approach to predict online purchasing intention. The methods involve conducting descriptive statistical analysis to discern variables correlated with shopping intention, providing valuable insights into factors influencing consumer behavior. The construction of prediction models, including Logistic Regression, Decision Tree, and Random Forest, demonstrates a comprehensive exploration of classical machine learning techniques. The evaluation of model performance using train accuracy and test accuracy adds a quantitative dimension to the study. However, the paper acknowledges limitations, particularly the exclusive consideration of aggregated page view data during the visit session and other user information. This limitation raises awareness of the potential oversight of other relevant variables that could significantly impact shopping intention, suggesting avenues for future research to enhance the predictive capabilities of machine learning models in this domain. I am planning to use VisitorType variable and utilize to predicting the purchase based on the model followed by the author.

Zhaoguang et al. (2021). In their paper “Potential buyer identification and purchase likelihood quantification by mining user-generated content on social media” [9] proposes a two-stage approach to identify potential buyers and quantify their purchase likelihood using user-generated content (UGC) on social media. The approach involves classifying user posts into before buying and after buying, and then using a novel Weighted Recency, Focus, and Sentiment (WRFS) model to quantify purchase likelihood. The method is verified using data from the Honda Civic community in the Bitauto automotive forum. The WRFS (Weighted Recency, Focus, and Sentiment) model is a novel approach used in this research paper to quantify the purchase likelihood of potential buyers based on user-generated content (UGC) on social media. The model considers three key factors: recency, focus, and sentiment. Recency refers to the timing of the user's posts, with more recent posts being given higher weightage. Focus refers to the content of the posts, distinguishing between posts made before buying and after buying. Sentiment refers to the sentiment expressed in the posts, such as positive or negative sentiment towards the product. By combining these factors, the WRFS model provides a quantification of the purchase likelihood for potential buyers identified through the classification of user posts. The accuracy of the model is demonstrated through the observation and verification of actual purchases made by these potential buyers. This paper is taking Recency (time), Focus (Informative page) and Sentiment (text based data) which are similar to my research question, they are too complex to this paper to explore.

What is the relationship between features related to timing (Weekend, Month, and special Day) and revenue generation?

Timing features such as weekends, months, and special days significantly influence revenue generation. Consumer behavior changes during weekends and special days like holidays or sales events, often leading to increased sales. Monthly revenue fluctuations can occur due to seasonal factors, such as increased retail sales during the holiday season or demand for certain services at specific times of the year. These timing features are crucial for revenue forecasting models, helping businesses anticipate revenue fluctuations and plan accordingly. However, the impact of these features can vary depending on the specific business and industry context.

Kawa et al. (2012) shows in the paper “A Descriptive Study on the Purchase Timing Effect in Athletic Shoes -Focused on Day-of-the-week Effect and Intra-month Effect” [10] describes the purchase timing behavior for athletic shoes, focusing on the day-of-the-week effect and the intra-month effect. It uses daily sales data from a domestic brand in Korea from January 2006 to December 2010. The results show that Saturday and Sunday have significantly higher sales than weekdays. Additionally, the first and third 10-days-of-the-month yield higher sales volume than the second 10-days-of-the-month. The department store's sales volume is higher in the first 10-days-of-the-month compared to discount and franchised stores, while the second 10-days-of-the-month have higher sales volume for discount and franchised stores based on the time series and statistical analysis.

In the “Data Analysis of Online Shopper’s Purchasing Intention Machine Learning for Prediction Analytics” paper Frazier et al. [4] proposes the use of basic web browsing analytics, machine learning, and artificial intelligence to effectively segment customers and identify those most likely to make a purchase, thereby improving the customer experience and facilitating purchases. It explores the impact of numerical and categorical features on the target class, using frequency, probability histograms, and scatterplots for numerical features, and frequency and probability analysis for categorical features. The paper provides insights into the relationship between weekend versus weekdays transactions and revenue creation, highlighting the importance of understanding user behavior and optimizing website design. It presents a Random Forest Classifier as an initial modeling framework to test the preliminary results of the data analysis [4].

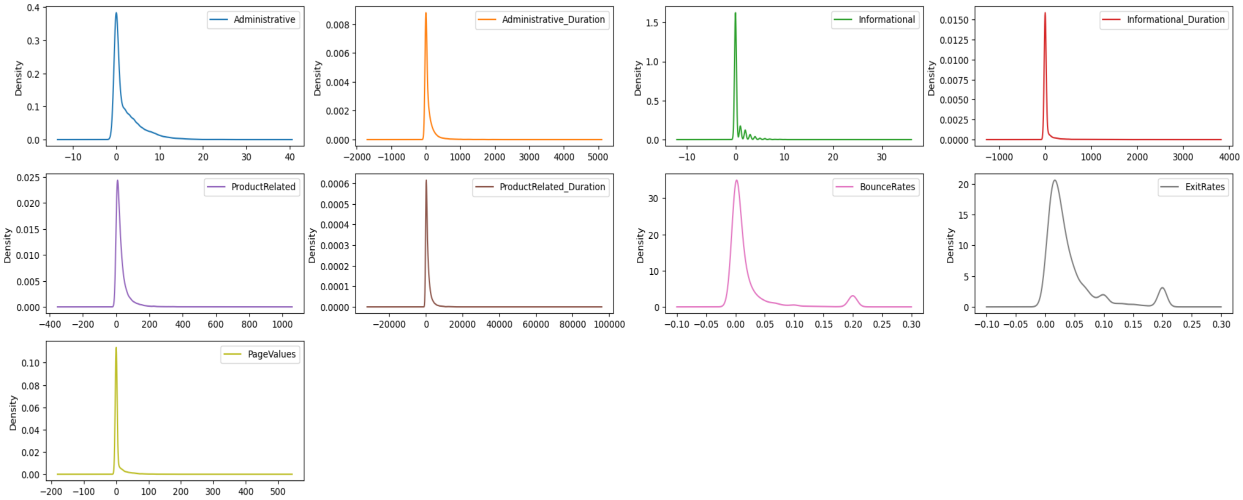
In my study, I am planning to use Linear regression, Decision tree, Random Forest as it is advantageous for my research question because it excels in handling non-linear relationships within timing features (Weekend, Month, Special Day) and their impact on revenue generation. Its ensemble learning, robustness to overfitting, and ability to handle missing values make it a powerful choice, providing valuable insights into feature importance for accurate predictions in the analysis.

Descriptive Statistics

Datasets: Online Shoppers Purchasing Intention Dataset

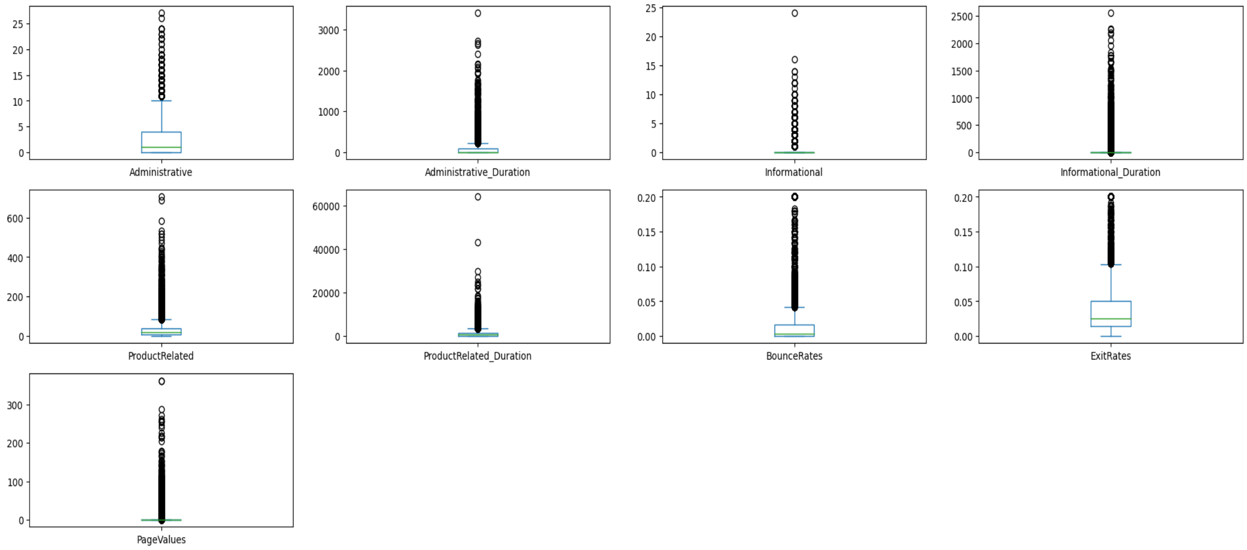
These datasets underwent preliminary cleaning, row data to technically correct data. Of the 12,330 sessions in the dataset, 84.5% (10,422) were negative class samples that did not end with shopping, and the rest (1908) were positive class samples ending with shopping. The dataset contains multivariate, with integer and real number. The dataset was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period. The profile reports will present detailed descriptive statistics about each column and its associated data. The tentative step by step Methodology from cleaning to the modeling is depicted below in the graph diagram below. Utilizing tools such as matplotlib and seaborn, EDA spans various critical steps, including data collection, summary statistics computation, data visualization through techniques like histograms and scatter plots, exploring data relationships, feature engineering, data transformation, dimensionality reduction, pattern recognition, hypothesis testing, and effective documentation and communication of findings. In the GitHub **‘eda_report.html’** is my final Exploratory Data Analysis.

Figure 1: density plots for numerical columns



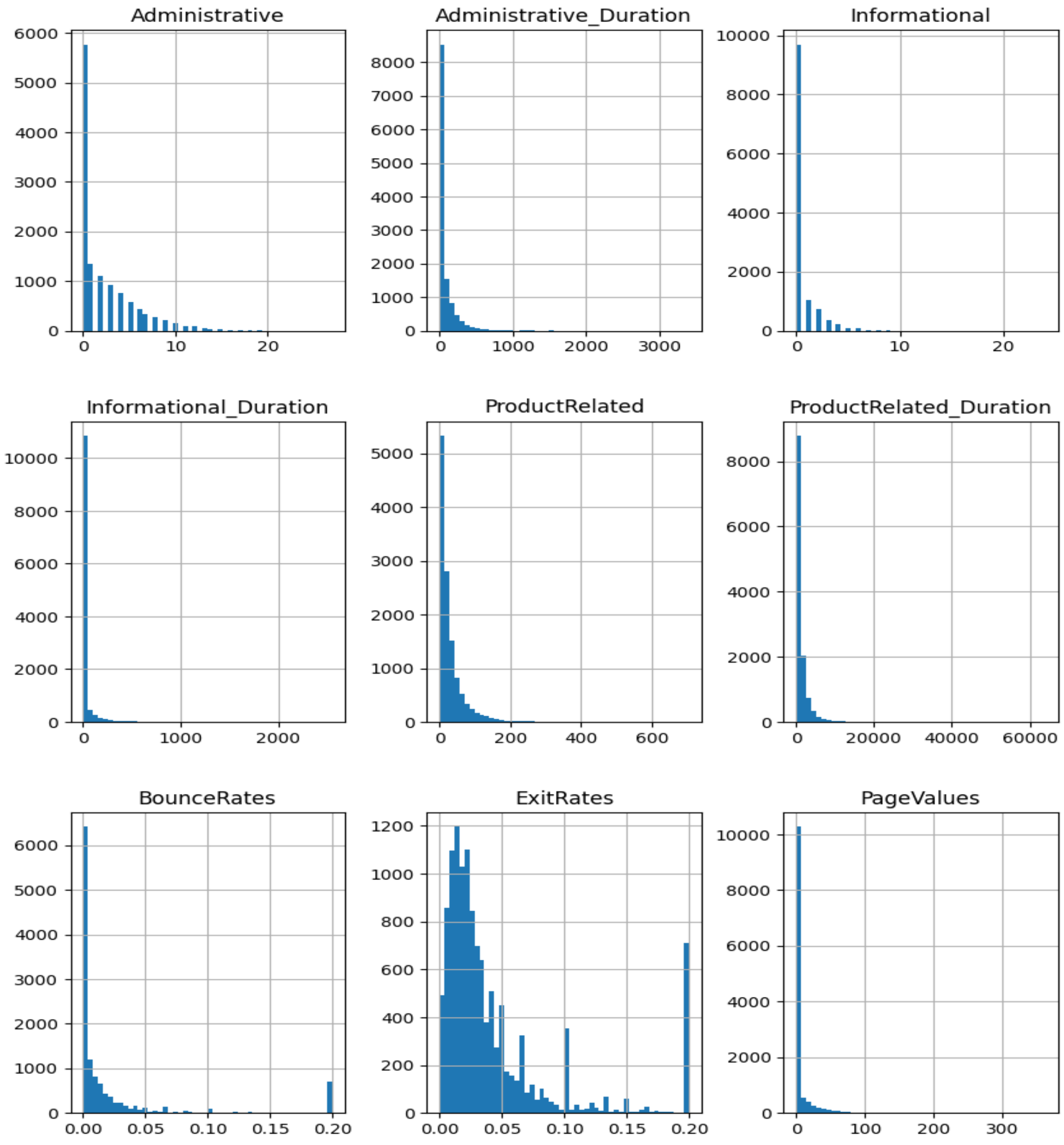
We observed that many numerical data points are skewed to the right, meaning a few users have very high usage numbers. This is typical in online shopping data.

Figure 2: box plot for numerical column.



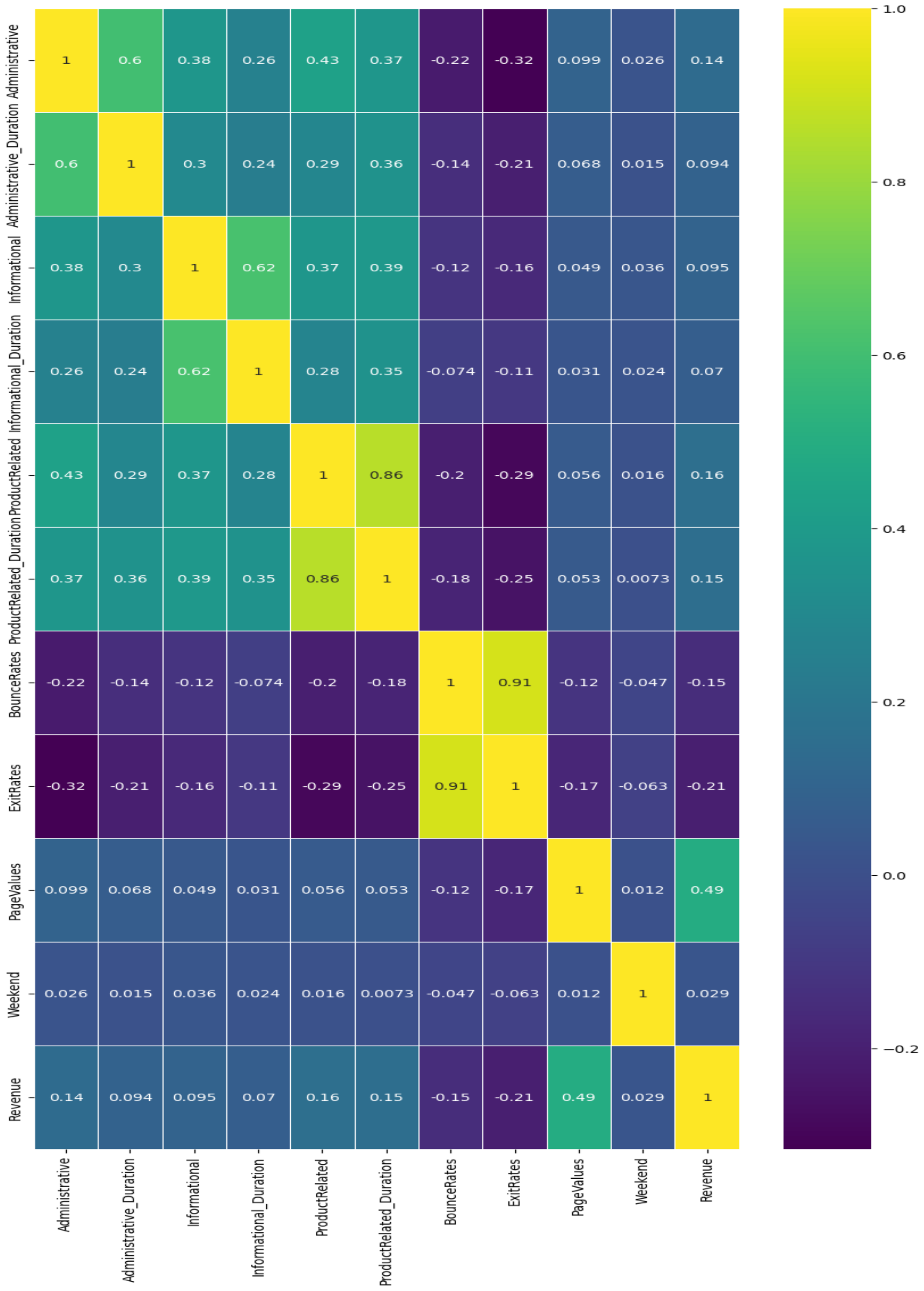
The graphical box plot provides a visual representation of the distribution of values in a numerical column. The box represents the interquartile range (IQR), with the central line inside the box representing the median. The "whiskers" extend from the box to the minimum and maximum values within a certain range (usually 1.5 times the IQR). Points outside the whiskers are considered outliers. The position of the box and whiskers gives an idea of the spread and right skewed of the data.

Figure 3: Histogram of numerical columns



Histogram shown in the figure explaines the shape of the data attributes as they are skewed to the right.

Figure 4: Heat plot of correlation



A correlation heatmap is a graphical tool that displays the correlation between multiple variables as a color-coded matrix. Understanding correlation heatmaps can help us identify patterns and relationships between multiple variables. The color of each cell represents the strength and direction of the correlation, with darker colors indicating stronger correlations. Positive correlations (when one variable increases, the other variable tends to increase) are usually represented by warm colors, such as red or orange. Negative correlations (when one variable increases, the other variable tends to decrease) are usually represented by cool colors, such as blue or green. We can see that product related duration and product related are highly correlated which make sense as these two variables represent the customers time spending on product related information. Bounce rate and Exit rate is also highly correlated.

GitHub Links

GitHub
https://github.com/KhaledSaiful/CIND-820-Big-Data-Project

EDA report:
https://github.com/KhaledSaiful/CIND-820-Big-Data-Project/blob/main/eda_report.html

Methodology

Literature Review

Literatures based on e-commerce transaction.
Literatures based on different machine learning algorithms.

Research Question

How do different informative page categories contribute to the likelihood of a user making a purchase?
Can we predict the likelihood of a user making a purchase based on visitor type (new, returning, other)?
What is the relationship between features related to timing (Weekend, Month, and special Day) and revenue generation?

Data Collection

Online Shoppers Purchasing Intention Dataset, this dataset was publicly available in 2018
Initial instances: 12,330, attributes: 18

Data Cleaning

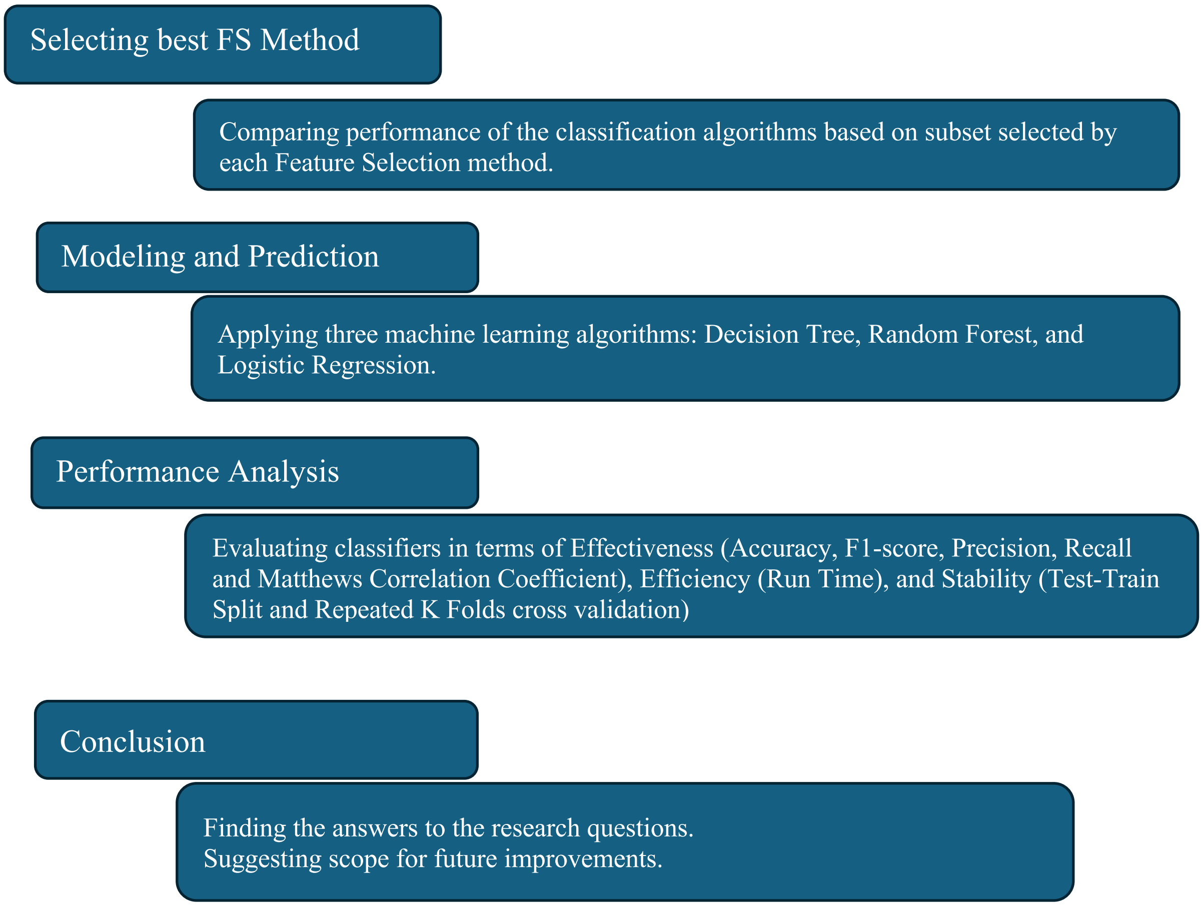
Check for missing values – no missing values.
Check duplicate values – contain duplicate values but ignored.
Check incorrect data type and convert to consistent data -Using Pandas astype function.

Exploratory Data Analysis

Univariate Analysis: Histograms and Bar graphs
Bivariate Analysis: Pearson and Spearman's Correlation
Using pandas profiling library/sweetviz for EDA.
Using Association rules and Visualization to gain insights to answer research questions.

Feature Engineering

Filter method
Wrapper Method
Embedded method



Selecting best FS Method

Comparing performance of the classification algorithms based on subset selected by each Feature Selection method.

Modeling and Prediction

Applying three machine learning algorithms: Decision Tree, Random Forest, and Logistic Regression.

Performance Analysis

Evaluating classifiers in terms of Effectiveness (Accuracy, F1-score, Precision, Recall and Matthews Correlation Coefficient), Efficiency (Run Time), and Stability (Test-Train Split and Repeated K Folds cross validation)

Conclusion

Finding the answers to the research questions.
Suggesting scope for future improvements.

References:

- Ading, Sunarto., Putri, Nilam, Kencana., Baliyah, Munadjat., Iriana, Kusuma, Dewi., Ali, Zaenal, Abidin., Robbi, Rahim. (2023). Application of Boosting Technique with C4.5 Algorithm to Reduce the Classification Error Rate in Online Shoppers Purchasing Intention. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, doi: 10.58346/jowua.2023.i2.001. [1]
- Baixue, Chen., Li, Li., Qixiang, Wang., Shun, Li. (2022). Promote or inhibit? Research on the transition of consumer potential purchase intention. Annals of Operations Research, doi: 10.1007/s10479-022-04777-2. [2]
- Kushwaha, Rahul., P., Anudeep., Desai, Shubham, Uday. (2023). Variable Aware Analytic Driven Online Shoppers Purchasing Intention using ML Algorithms. International Journal For Science Technology And Engineering, 11(5):5162-5168. doi: 10.22214/ijraset.2023.52878. [3]
- Andrew, Frazier., Fatbardha, Maloku., Xinzi, Li., Yichun, Chen., Yeji, Jung., Bahman, Zohuri. (2022). Data Analysis of Online Shopper's Purchasing Intention Machine Learning for Prediction Analytics. doi: 10.47363/jesmr/2022(3)162. [4]
- Rayhan, Kabir., Faisal, Bin, Ashraf., Rasif, Ajwad. (2019). Analysis of Different Predicting Model for Online Shoppers' Purchase Intention from Empirical Data. doi: 10.1109/ICCIT48885.2019.9038521. [5]

- Wenle, Wang., Jing, Wang., Yugen, Yi., Cui, Li. (2023). A User Purchase Behavior Prediction Method Based on XGBoost. Electronics, doi: 10.3390/electronics12092047. [6]
- Xiang, Shi. (2021). The Application of Machine Learning in Online Purchasing Intention Prediction.   doi: 10.1145/3469968.3469972. [7]
- International Journal of New Developments in Engineering and Society. (2021). Predict a UK Customer's Likelihood of Making an Online-purchase Based on the Logistic Regression Model. *International Journal of New Developments in Engineering and Society*, 6(1). doi: 10.25236/ijndes.2022.060104. [8]
- Zhaoguang, Xu., Yanzhong, Dang., Qianwen, Wang. (2021). Potential buyer identification and purchase likelihood quantification by mining user-generated content on social media. Expert Systems With Applications, 187:115899-. doi: 10.1016/J.ESWA.2021.115899.[9]
- Min, Ho, Lee., Sun-Jin, Hwang., Young-Sik, Kwak. (2012). A Descriptive Study on the Purchase Timing Effect in Athletic Shoes -Focused on Day-of-the-week Effect and Intra-month Effect-. Journal of the Korean Society of Clothing and Textiles, 36(4):422-431. doi: 10.5850/JKSCT.2012.36.4.422.[10]