Table of Contents

1. Abstract

The e-commerce company aims to enhance its operational performance and decision-making strategies by leveraging web analytics and predictive modeling techniques. With a dataset of 12,330 customer sessions, featuring browsing behavior and website interactions, the objective is to construct a predictive model that can determine whether a customer will make a purchase (Revenue = True) or not (Revenue = False). The analysis employs exploratory data analysis, clustering using K-Means [1], and predictive modeling techniques like Random Forest [7] [8] and Logistic Regression [9]. The models achieve testing accuracies of 89% and 87%, respectively, in predicting customer purchase behavior. Visualizations, model evaluation metrics, and key insights are provided to support decision-making strategies and improve the company's marketing efforts, resource allocation, and revenue generation. (Word count: 117)

2. Introduction

In the era of big data, organizations are increasingly leveraging web analytics and predictive modeling to gain insights into customer behavior patterns and optimize their operations. This case study focuses on an e-commerce company that has encountered challenges stemming from frequent breakdowns, unintended disruptions, and costly unexpected expenses, leading to losses and dissatisfied customers.

The company's website is a valuable source of data, and the organization aims to detect patterns and performance indicators that can inform effective decision-making strategies and improve operational results. The primary objectives of this study are:

1. Identify performance patterns and trends that will benefit the organization and its customers.
2. Provide effective decision-making strategies for the company.
3. Forecast and predict unsolicited events, such as failures or required methods.
4. Construct a predictive model using customer visit data from the website.

For meeting the predetermined targets, it will capitalize on the following dataset: 12,330 customer session visits to the company's website, consisting of several types of sources including administrative, informational, product-related, bounce rates, exit rates, page values and special days. Exploratory data analysis, clustering tecniques Random Forest and Logistic Regression used for prediction modeling will be occupied to find out answers and build predictive model.

We will do an in-depth review of the data and create all necessary visualizations to build a solid ground for taking the smartest decisions possible. Thus, in addition to the predictions tabulated in the feed performance charts data will also be provided to support improvements in marketing efforts, resource allocation, and revenue generation.

3. Literature Review

An important part of the analysis of a website is in collecting, reporting, analyzing, and evaluating the website data in terms their effectiveness in meeting the organizational goals. Major metrics tracked include the number of visitors, their traffic sources, page views, dwell times, and conversions The predictive web analytics further move ahead by applying such methods as machine learning in order to forecast the further customer behavior by getting access to past database. It thus drives actions like personalized marketing, customized content suggestion, and goal-oriented campaigns that are return-driven.

This case study utilizes a data set with 12,330 visits of customers session to the platform website [12]. It was designed so that every session considering whether it has a landing day, campaign, user profile, or a period would be smoothened for the past year. The dataset is formed by a ser of ten numeric and eight categorical attributes. It includes they:

- Revenue: The target variable indicating whether a purchase was made (True or False).
- Administrative and Informational: Number of administrative and informational pages visited, and the total time spent on each category.
- Product Related: Number of product-related pages visited and the total time spent on them.
- Bounce Rate: Percentage of customers who exit a site when landed on a particular page.
- Exit Rate: Percents of the outgoing on a special page.
- Page Value: Visited before transaction completion, average value of an previously visited web page before complete.
- Special Day: Closeness of the visit to a specific special day.

The dataset also includes attributes such as browser, operating system, region, traffic type, visitor type (returning or new), and Boolean values indicating whether the visit occurred during a weekend or a specific month.

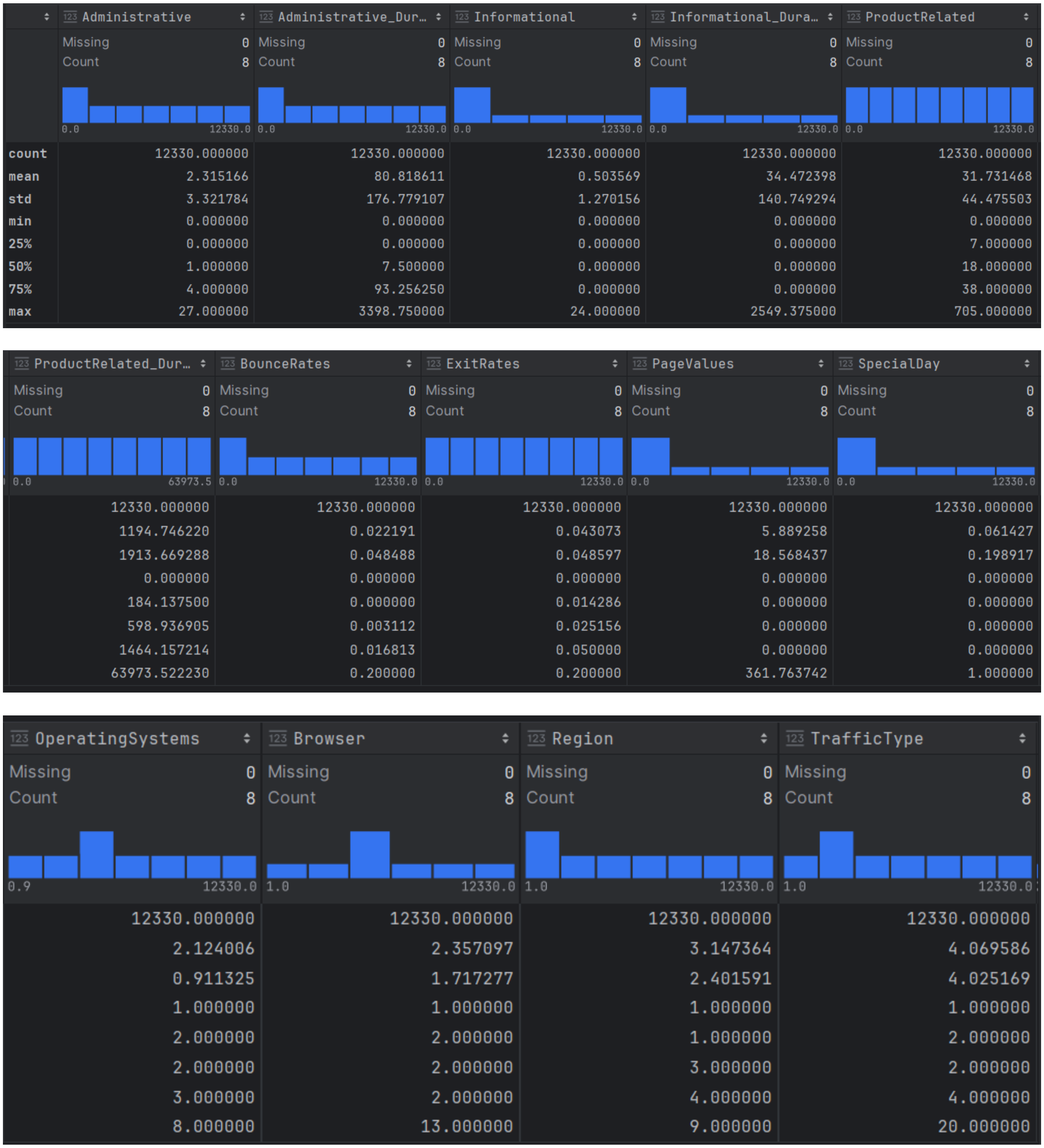
The architecture of the analysis involves exploratory data analysis, clustering techniques (e.g., K-Means), and predictive modeling methods like Random Forest and Logistic Regression. These techniques will be applied to uncover patterns, segment customers, and construct predictive models to forecast customer purchase behavior, supporting decision-making strategies and operational improvements for the e-commerce company.

4. Discussion and Analysis

4.1 Data Analysis

As per the traditional approach, we first shall observe the structure of our data. Basically, we shall learn what our data is and how it looks like.

BELOW IS THE SUMMARY STATISITICS OF OUR DATA



	Administrative	Administrative_Dur…	Informational	Informational_Dura…	ProductRelated
	Missing 0	Missing 0	Missing 0	Missing 0	Missing 0
	Count 8	Count 8	Count 8	Count 8	Count 8
count	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000
mean	2.315166	80.818611	0.503569	34.472398	31.731468
std	3.321784	176.779107	1.270156	140.749294	44.475503
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	7.000000
50%	1.000000	7.500000	0.000000	0.000000	18.000000
75%	4.000000	93.256250	0.000000	0.000000	38.000000
max	27.000000	3398.750000	24.000000	2549.375000	705.000000

ProductRelated_Dur…	BounceRates	ExitRates	PageValues	SpecialDay
Missing 0	Missing 0	Missing 0	Missing 0	Missing 0
Count 8	Count 8	Count 8	Count 8	Count 8
12330.000000	12330.000000	12330.000000	12330.000000	12330.000000
1194.746220	0.022191	0.043073	5.889258	0.061427
1913.669288	0.048488	0.048597	18.568437	0.198917
0.000000	0.000000	0.000000	0.000000	0.000000
184.137500	0.000000	0.014286	0.000000	0.000000
598.936905	0.003112	0.025156	0.000000	0.000000
1464.157214	0.016813	0.050000	0.000000	0.000000
63973.522230	0.200000	0.200000	361.763742	1.000000

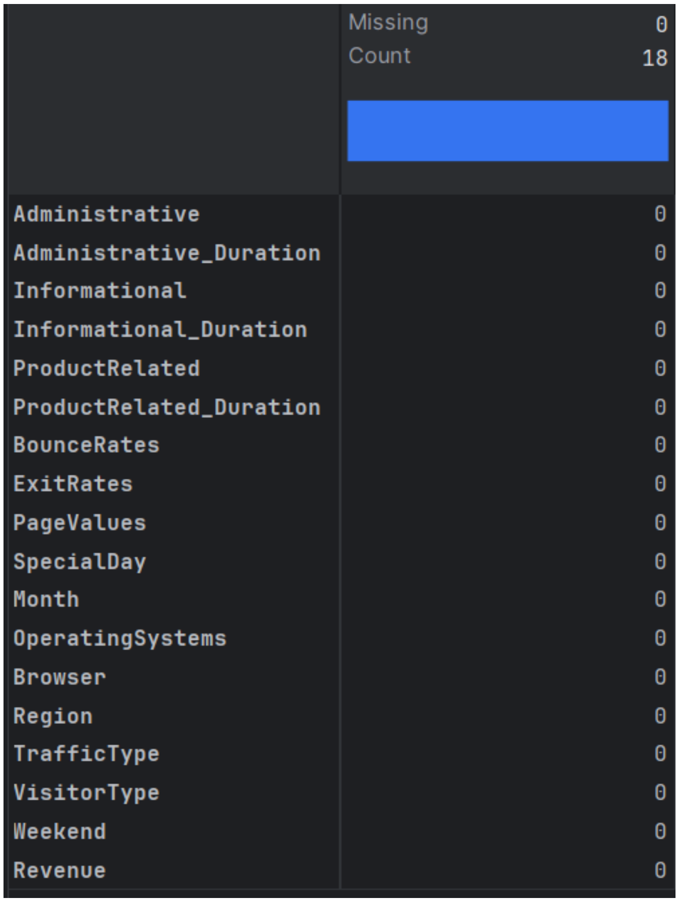
OperatingSystems	Browser	Region	TrafficType
Missing 0	Missing 0	Missing 0	Missing 0
Count 8	Count 8	Count 8	Count 8
12330.000000	12330.000000	12330.000000	12330.000000
2.124006	2.357097	3.147364	4.069586
0.911325	1.717277	2.401591	4.025169
1.000000	1.000000	1.000000	1.000000
2.000000	2.000000	1.000000	2.000000
2.000000	2.000000	3.000000	2.000000
3.000000	2.000000	4.000000	4.000000
8.000000	13.000000	9.000000	20.000000

fig1: summary statistics

The dataset includes ten numerical attributes and eight categorical attributes. The "Revenue" attribute indicates whether a transaction was made, with possible values being False or True. Attributes such as "Administrative" and "Administrative Duration" represent the number of administrative pages visited and the total time spent on them per session, while "Informational" and "Informational Duration" represent similar metrics for informational pages. "Product Related" and "Product Related Duration" describe the pages related to products visited and the time spent on them. "Bounce Rate" indicates the percentage of visitors who leave the site after viewing a single page without further interaction, while "Exit Rate" shows the percentage of exits on a page. "Page Value" represents the average value of a web page visited before completing an e-commerce transaction. The "Special Day" attribute indicates how close the visit is to a specific special day.

Additionally, the dataset includes attributes such as browser, operating system, region, traffic type, visitor type (returning or new), and a Boolean value indicating whether the visit occurred during a weekend or a specific month. Some of these attributes do not seem to contribute anything in our predictive as per common sense and it shall not be wrong to remove them.

Similarly, let us check for any missing or bad values in our dataset.



	Missing 0 Count 18
Administrative	0
Administrative_Duration	0
Informational	0
Informational_Duration	0
ProductRelated	0
ProductRelated_Duration	0
BounceRates	0
ExitRates	0
PageValues	0
SpecialDay	0
Month	0
OperatingSystems	0
Browser	0
Region	0
TrafficType	0
VisitorType	0
Weekend	0
Revenue	0

fig2: missing values

phew! No missing values found!

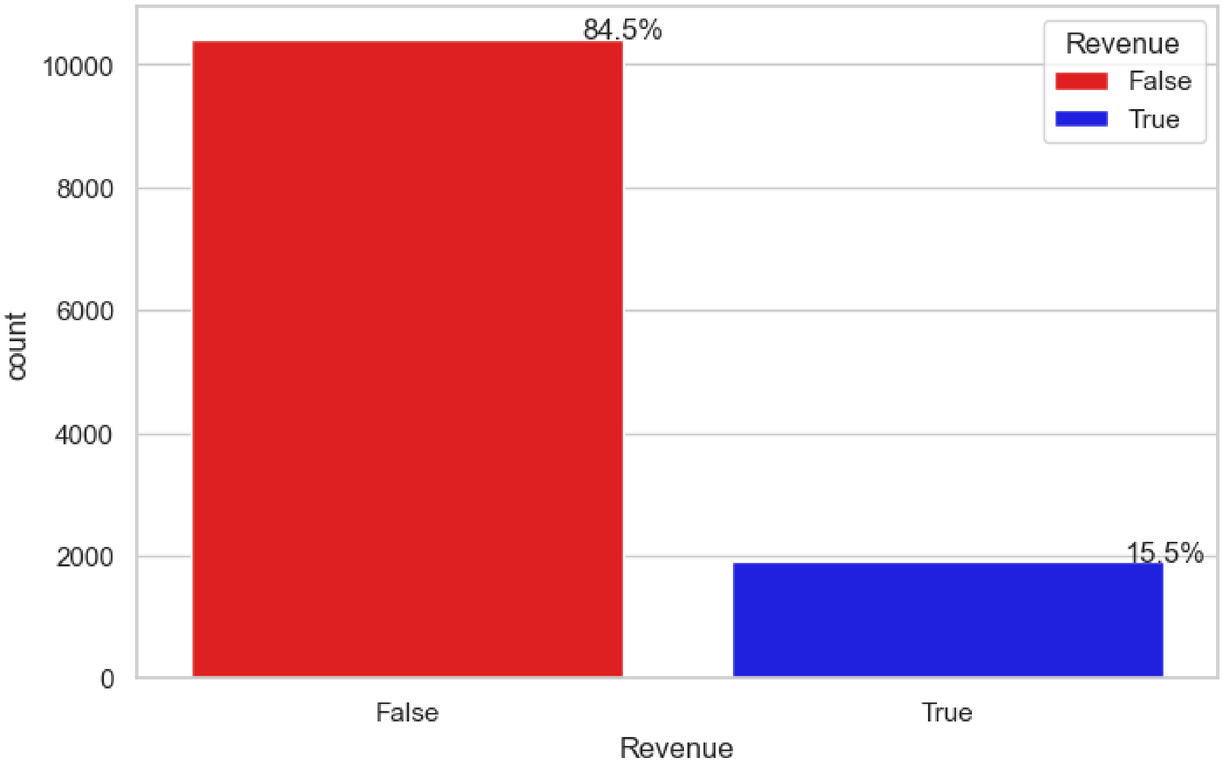


fig3: revenue distribution

Many of the entries were captioned as "False" on the "Revenue" column, which sums up around 85.5% of the entries. On the contrary, things are quite different when it comes to the estimation of true revenue, where just a half of the cases are correctly identified as "True" for the purpose of revenue generation. That the class imbalance affects predictive modeling, and this can cause bias towards non-purchase results which will result to inefficiency in organizing marketing plans [1]. Therefore, for the purpose of overcoming this dilemma, ideas such as oversampling, undersampling or the ensemble methods can be incorporated to the model during the training. On the other hand, precision may not be the only competent criterion for critiquing model performance by virtue of lopsidedness in the data. On the contrary, precision, recall, F1-score, or AUC-ROC [17] should be emphasized for a full review of performance, especially for the instance of distinguishing "buy" from "no-buy" instances.

Disparity in the propagation of the "Revenue" class attribute with 85.5% and 15.5% summed up as "False" and "True" respectively shows the organization's difficulty in pushing sales from just the website. It is highly important to develop an action plan aimed to analyze the implications causing this type of disparity as a necessity to optimize production success and increase revenues. If the organization can cross-reference and analyze client visit data with predictive modeling methods, it can forecast revenue potential, identify high-value customer segments, and target its marketing more precisely so as to maximize the revenue generation opportunities. Data interpretation of the App's revenue class data can fuel decision-making processes, reflecting in activity expenses, marketing strategic planning, and website fine-tuning. Addressing the underlying factors that contribute to the revenue gap is key to the business being able to work on deeper strategies that will ensure both short-term and long-term revenue growth and, hence, provide better operating performance.

One for all to look at is how new visitors react to your website as well as those that are regular customers.

.

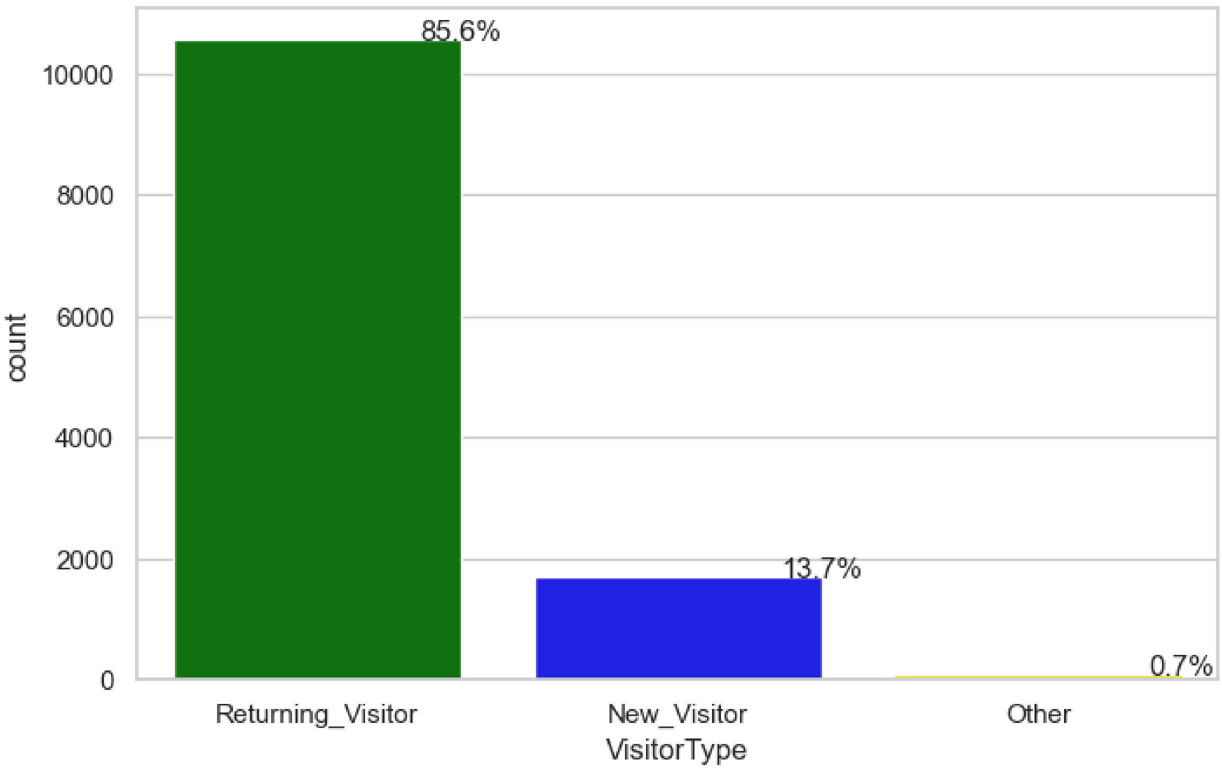


fig4: customer visiting types.

In the category of Visitor Type, the dataset exhibits that more 85.6% of total visitors to the website fall under the ReturningVisitors category. Conversely, the second segment (NewVisitors) amounts to 13.7% that remains about 0.7% of the dataset that is included as Others. The trends in the display indicate the high number of returning customers as opposed to the new customers' interest acquisition.

The most important aspect of Customers is a customer who has a habit of repeating this behavior. It indicates a strong customer base and also showcases the need to take care of these sorts of relationships which relate to a repeat business and ultimately results in the long term customer value. For example, strategic tactics like personalized recommendations, customer loyalty programs, and targeted offers can be implemented to inspire regular returnees and continually involve them in the webpage.

Such us, the emergence of New Visitors recommends the possibility of customer acquisition and hidden potential in growing the customer base. Through analyzing emerging customers' behavior patterns and way of thinking, the organization can personalize onboarding experience, amplify conversion rate, and raise the efficiency of target campaigns to eventually close more deals.

The overview of Others might not be comparable but it's still effective because it may include unusual visitors and market segment that can be studied deeply. Knowing the specific elements of the groups of visitors to the region offers a chance to detect the current gaps, as well as the further chances for increasing the amount of the customers who are satisfied.

Hence this summarizes the overall, why don't we divide it to weekly behavior or weekends especially?

Analyzing customer visit types analysis by week makes possible to leverage time series analysis and make many interesting time dependence features to be revealed that cannot be seen by traditional methods of long periods viewing [13]. Weekly analytics are the fastest mode for a more detailed comprehension of visitors, preferences, and leading behaviors in smaller time intervals [14]. Through analyzing visitor types on the weekly basis, the businesses will detect recurring patterns, including days of the week, or times when the business receives more visitors [15]. These will be the main points for effective strategies regarding use of resources, marketing campaigns and website optimization. Besides, weekly analysis is also a key factor that makes a difference due to short-notice adjustment of marketing strategies and promotions based on the observed trends, which is the variety particularly in the consumer behaviour. In addition, monitoring daily visitor types on a weekly basis is useful for the measurement of the effectiveness of any ongoing marketing or promotional activities targeting particular market segments that are conducted within a particular week and evaluation of their effect on different visitor groups. Elaborate, visit type analysis by week will give managerial staff insights to deeper understanding of visitor behaviour [16]. This way data driven decisions will be possible and changes to strategies will be done as best as possible.

.

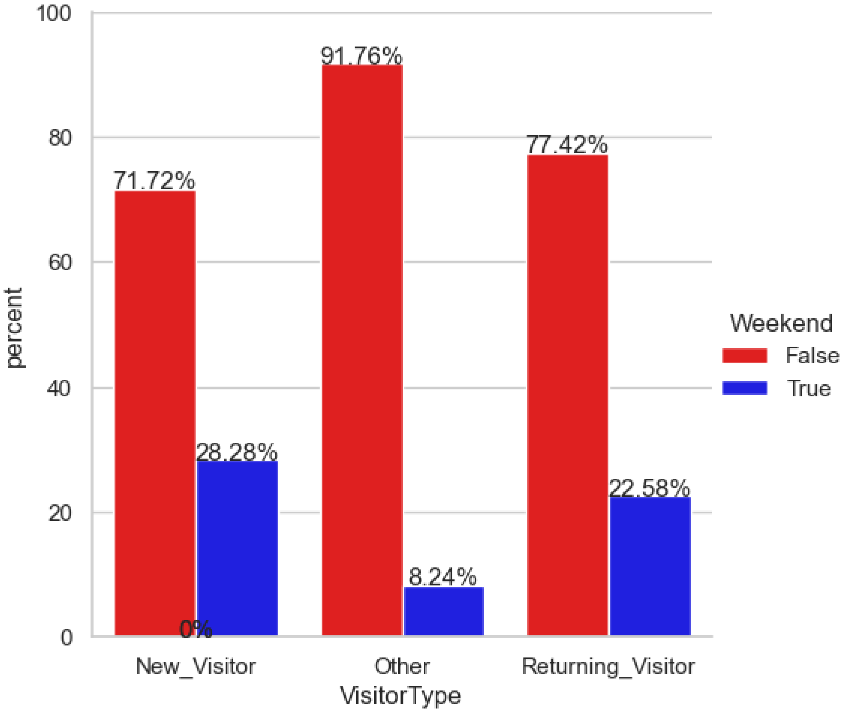


fig5: customer weekend & weekdays visits

New Visitors: This unevenness among the new visitors that 71% of them did not come on weekends while 29% of them did come on weekends shows a sizable number of remarkers among this group. This finding can be used as the basis of campaign creation for weekends or promotions for people that have always wanted to visit; this can be utilized on their weekend scrolls to facilitate conversion of browsers into buyers.

Other Visitors: Another factor to consider is that of other visitors whereas you represent only a small percentage of the total visitors (0.7%). This is however noteworthy because 91% did not visit the park on weekends, the ‘other visitors’ segment and 9% on the other hand could only visit the park during weekends. This means that understanding the characters and likes of each visitor category helped in knowing what to adjust for better Some visitors seem to be concerned with issues only during the weekends, which is when they are free. Using such information can, therefore, help you to draw visitors closer that may convert to consumers of your services.

Returning Visitors: The result which shows that 77% of visitors who returned again during working days and 23% of active on weekends illustrates the necessity to increase liaising with returning customers during all the days of the week. Methods like personalized offers, a land promotion for the weekend only, or unique content can make visitors coming back to the place want to visit again and in fact, make them come back for the second time. Thereby, these customers will be retained and consequently, will drive return sales.

We have observed the customer trends from fast consumption to slow sales. Sunday is the quietest day while Friday is the busiest. Now look at the customer data using the approach of break out with monthly and geographical means.

Being able to map customer distributions per months not only provides vital information for seasonal patterns, cyclically trends, and temporal variations of customer behavior, but also assists in creating effective marketing campaigns, efficient inventory management, and appropriate decision making related to the operational planning. Through an analysis concentrating on the months where it can find that the customers number is higher with some specific months that happens to be high demand, their period of engagement or seasonal fluctuations when they can adjust to their strategies and management on resources that are specific. On illustration, shop owners can use the data to ensure that they plan for the entertainment programs aligned to consumer behavior, preferences, and seasons. On the one hand, industries whose businesses revolve around services can forecast drops in demand for their services and accordingly, they can reduce or even change their staffing levels or their services to tackle customers' demands. Besides that, illustrating customer distributions based on Months makes for organization to understand how their marketing campaigns, new products or seasonal promotions has done since these periods and to apply money based strategy and also improve the marketing strategy of them. In general, by knowing their customers-month the organizations can seize the seasonal chances, avoid the seasonal troubles and offer the high-class customer services every time, in the long run they can impress, grow the business and succeed.

.

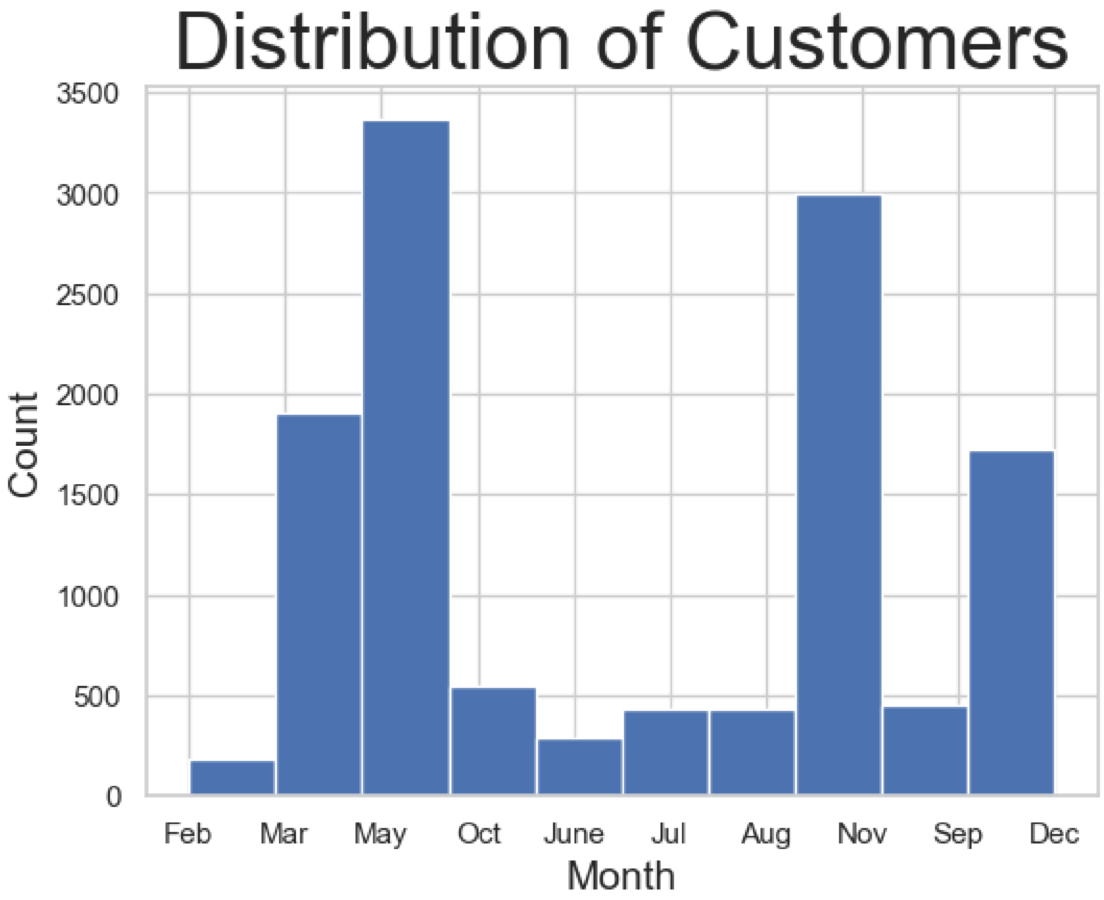


fig6: monthly behavior of customers visits

A distinct correlation is seen between months and customer distribution patterns as the analysis offers useful conclusions that can help the business not only from operational point of view but also in the development of other related issues.

Balancing the need for advancements with safeguards for human values and psychological well-being presents challenging ethical dilemmas for society and policymakers.

1. **Seasonal Demand Planning**: The fact that the biggest customers' count is in May, November, and December is indicative of intense or peak seasons in those times. Organizations could employ this information to help fine-tune procurement calls, staffing schedules and resources levels to reduce wastage in order to accommodate the increased demand during the peak season. For instance, you may apply adjustments to the stock levels as well as the staffing to the plans since this is the time the customers need you most. Also, if your enterprise is a service-oriented one, you may adjust the service that you are offering and the capacity to fit in the anticipated size of the customers during the offseason.

2. **Marketing and Promotion Strategies**: Learning about seasonal patterns in numbers of tourists/guests shall push previous organizations to customize their marketing and promo strategies so that during peak periods they can be effective and engage customers. One illustration is that a business could carry out a planned marketing campaign targeted at peak months, special offers, or seasonal promotions that run during the remarkable months like May, November, and December to enable it effectively to achieve its customers retention goals. Furthermore, by way of using customer distribution data, businesses could take advantage of it to know the best time and the message to send to the customer to make the marketing communication more relevant and effective.
3. **Operational Efficiency**: Providing a dynamic approach to demand by using the trend identified, companies to enhance the efficiency and quality of the service delivery by working in accordance with the observed patterns. We see that enterprises can run their processes smoother: e.g., enhance order fulfillment, customer support service, and supply chain operations without disappointing clients.

NOW FOR REGIONAL TRENDS

A recurring task that provides in-depth information about customers' geographical tendencies, inclinations, and behaviors is visualization of customer distributions by region, which enables to strategize actions, plan resources, and make operational decisions based on the newly received knowledge from the data [18]. The internalization of the client's distribution by the region enables businesses to recognize the presence of customers in the highest density regions or markets that are emerging, though some regions remain underserved. This allows the businesses to tailor their products, services, and marketing to suit the specific needs and likes of the geographic area. Besides, geographic visualization of networks of the customers allows companies to recognize geographic clusters or tendencies which affect customer behavior, like cultural differences, socioeconomic status, or special market conditions [19]. This insight will, thus, contribute to producing decisions that are strategic and regionally sensitive in nature, and will be focused on customer engagement, satisfaction, and retention. On the other side, tracking fluctuations in customer distribution in each region can assist organizations to change with times and implement new strategies based on market conditions that are changing [20]. Also, it can strengthen performance by maximizing new opportunities and addressing challenges early.

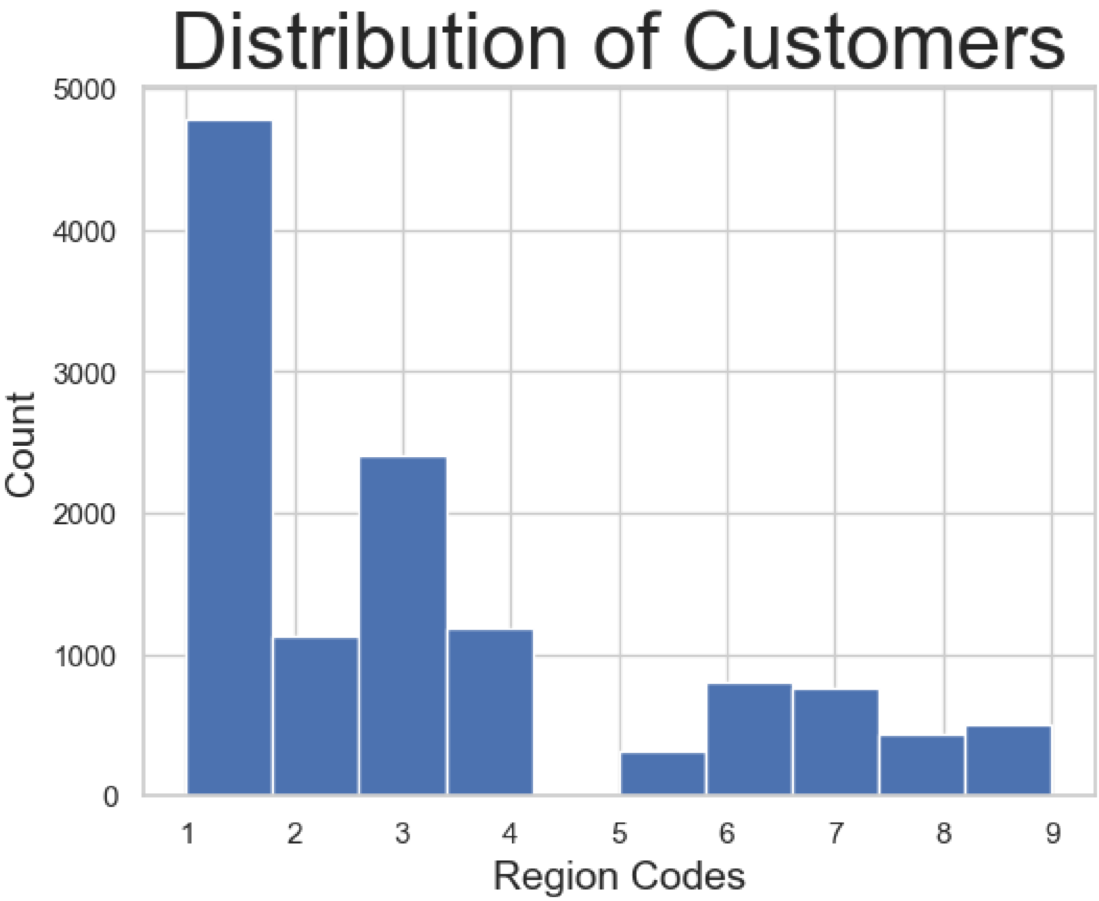


fig7: regional distribution of customers

We have the representation of distributed customers across various regions depicted in fig. 7. polarized from the consumer's perception and generates key details that can be utilized in the formulation of multiple decision processes in the organization.

1. **Regional Targeting**: The above high market share in region 1 indicates either a high rate of market penetration or customer loyalty in that region. Organizations may then use this knowledge to channel their resources, marketing strategies, and customer interaction efforts more competitively in Region 1, to take advantage of the market potential that already exists and cause even faster growth to occur. On the other hand, one of the important things which can be profoundly important while considering factors driving the customer attraction and retention in region 1 is how this information can be useful for expanding into the similar other markets or replicating the successes in other regions.

2. **Market Expansion Opportunities**: Although Region 1 could be a large market opportunity, the low customer count in region 2, Region 3, and Region 4, this implies that the concept of market expansion and business growth through potential opportunities is mainly evident in this regions. Organizations can embark on such market research information and then be able to identify potential zones which still have not been tapped on and newly emerging markets pinpointing the best strategies to enter and prosper in them. If a company is ready to spend on conducting market studies, customer acquisition efforts and locally targeted marketing strategies then they have a chance to explore all growth avenues these regions offer and thus they would have a more diverse customer base.
3. **Customer Segmentation and Targeting**: Analyzing consumer distribution by region allows organizations to spot clusters or meaningful spatial patterns of demand which in turn explain to customer behavior and attitudes. One of the ways in which customer segmentation [4] helps organizations to achieve effective marketing is through the placement of customers in geographic zones. By doing this, organizations can craft marketing messages, product offerings, and promotional activities that are more meaningful and desirable to the unique needs and preferences of every region. This customer segmentation and targeting approach that is pinpointed to the needs of customers can add value to customer engagement, spike up conversion rates, and bring marketing to the forefront [4].
4. **Operational Efficiency**: Increased knowledge of the geographic distribution of clients can also inform the other operational decisions involving logistics, supply chain management and networks. Through distribution of customers by region, the companies can utilize distribution channels, inventory management strategies and operational procedures to cater for the needs of customers in each region at the lowest costs and maximum efficiency.

Visualizing revenue with respect to page values in a strip plot

Getting into details of revenues against the page values with the stripe-plot diagram provides the valuable knowledge concerning the connection between customer related engagement and revenue generation, providing the ground for data-based decision making and success in the optimization of marketing strategies. Through putting the page rank and returns on other axis, organizations can do visualizations to see the relationship between the two variables and find out trends. This way, we could understand what pages actually help to turn revenue the most and what pages bring less pageviews to be optimized or studied further. Moreover, marketers were able to access information on revenues by segmenting the data according to revenue outcomes (True for received revenues and False for the no revenues generated). Thus they got to understand pages' effectiveness in driving revenues and further modified their marketing efforts, content strategies, and website optimization initiatives accordingly. In conclusion, pictorial representation of revenue in strip plot equips businesses to profit more by way of customer engagement, revenue generation and figure-out their digital marketing efforts effectiveness.

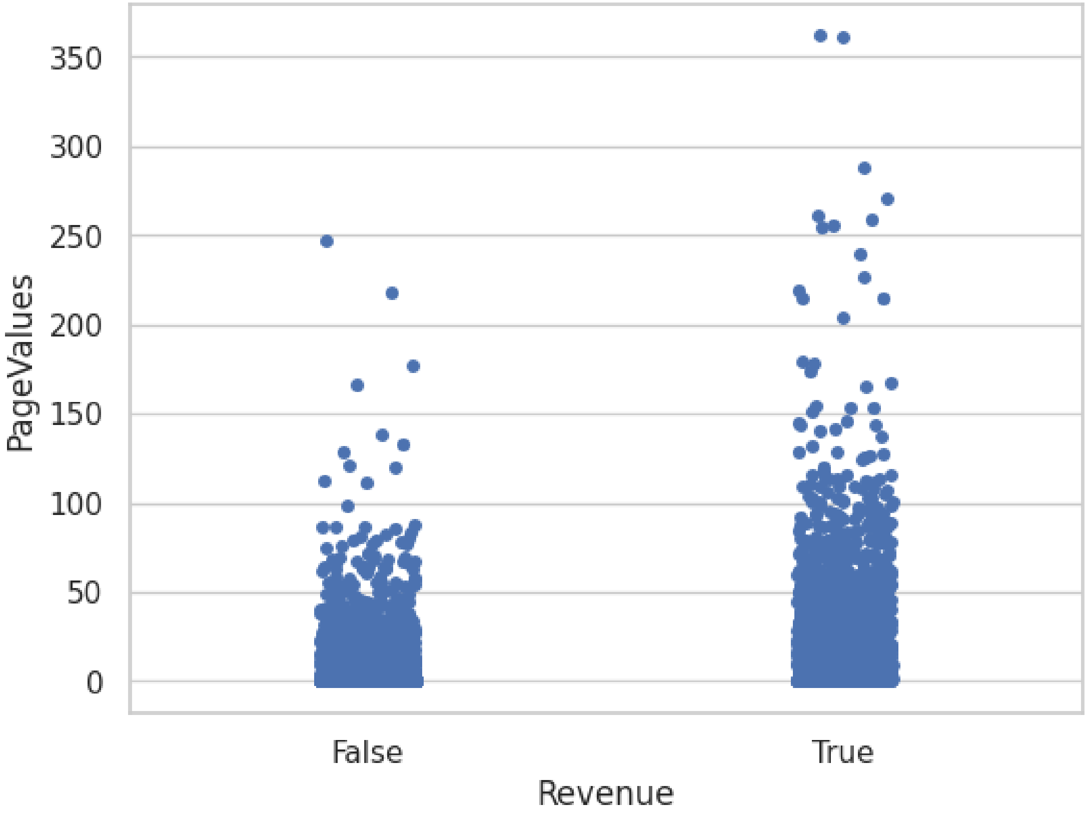


fig8: strip plot of revenue with respect to the page values

The detected image on the radar plot which showed the low values page next to high revenue in both category (true and false) might be weird to first sight. May be true to say so, but mind must take this possibility and vexatious elements that could emerge into consideration as well as the effects of such practice on the decision making.

1. **Landing Pages and Conversion Optimization**: A low credit value might represent the money spent on your business which you received after the session, and this could be seen as revenue from the conversion of leads without any visits to the site which later improved the value of the page. Eg, these pages may have been intended to bring in customers with certain preferences. It thus avoids sales and good revenue generation. Organizations can dissect the content, look, and user-friendliness of such pages to find out underlying causes responsible for their fame and use represented approaches across other website's pages as well.

2. **Transactional vs. Informational Content**: Pages mentioned in this category typically involve transactional matters dealing with products, checkout processes and similar content that generates revenue for the business in the first place. On the contrary, the more significant values are usually for pages that contain informational or supplementary content that does the work of supporting the audience process but doesn't necessarily cause the conversions. Knowing which types of web pages are important in the conversion funnel and how to optimize the web page's structure and content to get rightful and useful revenues is very fundamental.

3. **Outlier Detection and Data Quality**: A presence of outliers in higher page values may billow at anomalies and data irregularities that need a second thought to gather evidence on them. Organizations need to relate the information and detect outliers during the data validation and reliability process for data accuracy and reliability. Another important aspect of the analysis is that it can bring out invaluable information as concerning clients' behavior, and find out the emerging trends, and help in making decisions about the marketing strategies, contents of the website, and website performance.

4. The apparent inconsistencies on the website data can be reconciled when it is evaluated thoroughly and viewed as a whole. Insights from different data such as visitors' flow, decisions made by visitors, etc. should be considered. Through determining such correlation, firms can single out the chances of optimization, give proper directional contribution to the efforts of marketing and content, and, overall, increase the profitability of the organization. Other features such as outlier detection and data quality assurance are needed for ensuring the helpfulness of results gotten through visual exposure of strip charts.

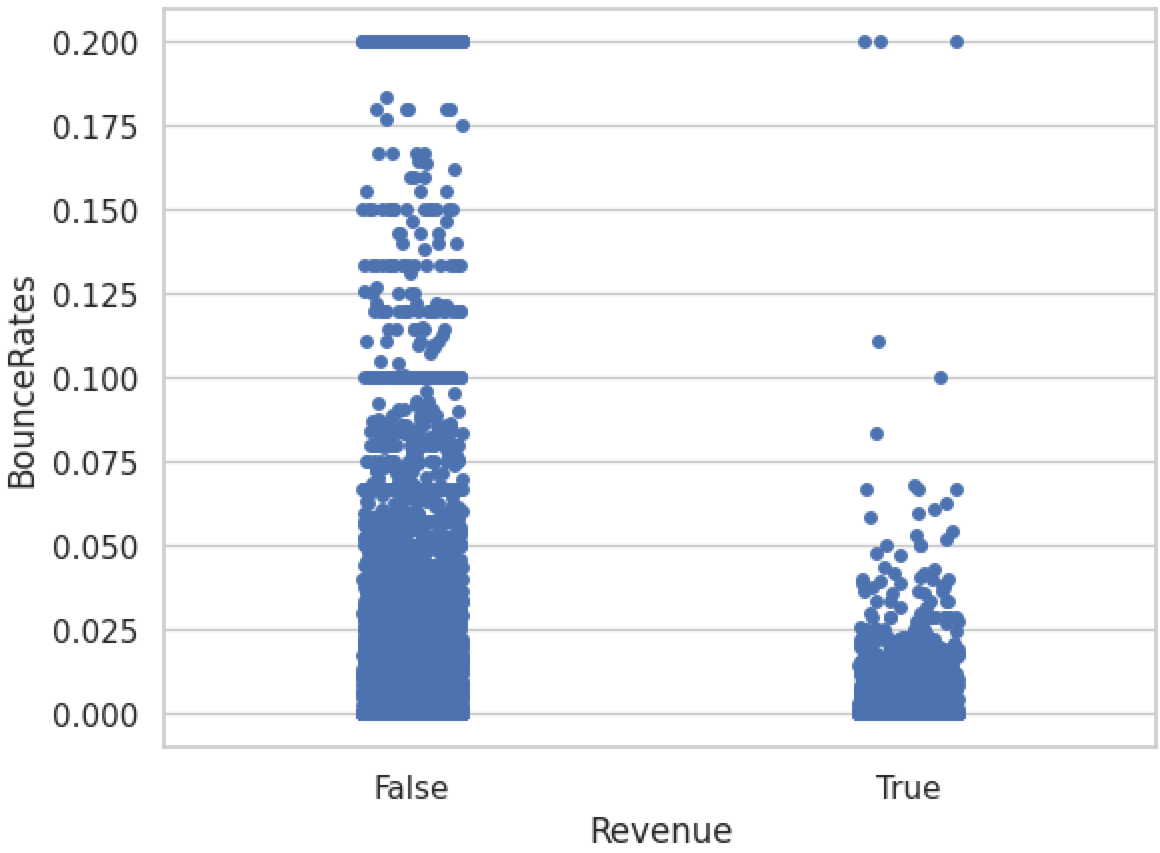


fig8: strip plot of revenue with respect to the bounce rates

The visualization of the strip plot indicates that there is a certain divide between the rate in which the customers bounce from the site and the clusters that represent their spending outcomes, which offers a useful perspective of the connection between bounce rate and revenue.

1. **Optimizing Bounce Rates for Revenue Generation**: The tight data grouping for 0-0.100 screent bytes in the False revenue outcome means that low bounce rates generally result in higher revenue production. The above allows us to understand the purpose of designing a website, and come up with high quality content that will keep customers on your site longer so that they can engage with your site. For a company, it pays to focus on landing page quality, useful content, and customer-friendly navigation menu. This measures the number of people that quit your website and can turn potential leads into your loyal buyers.
2. **Identifying Outliers and Anomalies**: The existence of the outliers in the buckets with the high bounce rates(for example,0.200) for the True ringing results indicates that there might be some anomalies or unusual routines in the behavior of visitors. It is normal to have some

degree of fast exit especially if visitors find what they are looking for immediately, but abnormally high bounce rate may be as a result of issues on website usability, content quality or reaching the wrong target market. Organizations need to study the appearance of such cases and pinpoints whether their site visitors are not being engaged or converted and the implementation of the remedy is highly needed to enhance the performance of the site and even the conversion rates.

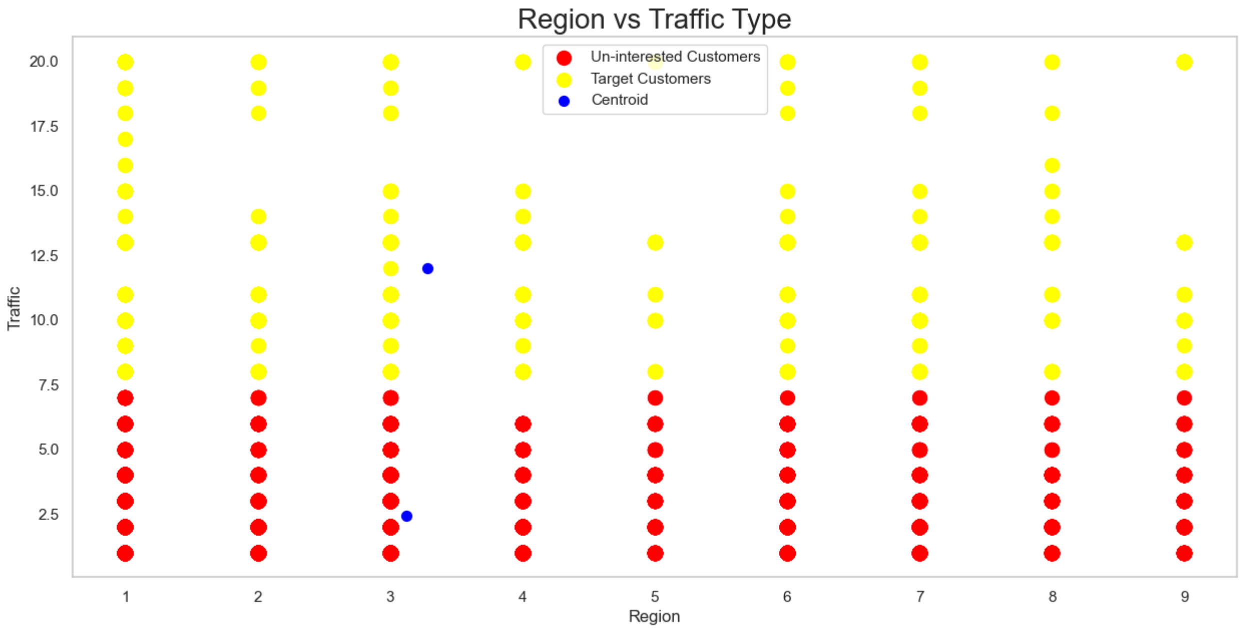
3. **Segmentation and Targeting**: Bounce rate-revenue analysis of visitors versus their engagement levels then enables organizations to divvy up users between those who engage more and, tailor-made strategies and marketing for each of them. The example given poor hits may suggest the readiness for management, whereas ones who stay on the website for a long time are certainly not a focus on conversion. Education of the connection between bounce rates and revenue allows organization to improve their marketing techniques and content strategies as well as make their website experience tailored to meet the needs of audiences in a way that maximizes revenue and therefore the overall business success.

The conclusions observed via the strip plot form a basis for drawing conclusions about the bounce rate-to-income trends that accompany web design, visitor engagement, and revenue. These outcomes help to inform decisions related to optimizing websites, user engagement, and revenue generation. Through enhancing bounce rates, locating outliers, and using segmentation opportunities companies will discover the level of effectiveness of their digital marketing which will drive sustainable income growth.

CLUSTER ANALYSIS

By using cluster analysis [3] integration in the case study it is possible to reveal untold patterns and segments in the client base data, as a result of which it is possible to develop a targeted marketing and to make personalized customer experience. Through customer segmentation [4] which is based on their view history, spending habits, and social media data, we can find up distinct customer groups that vary in their tastes, behavior, and preferences. With the help of this segmentation, we can customize the marketing campaigns, products and website experiences to the perfection for particular customer groups, which positively affect the customer satisfaction level, retention rate and profitability of these businesses.

The course of a traffic scenario by region helps scientists analyse data on the differences in geographic behavioral patterns and customer preferences. Visualizing how the traffic mix (like direct, organic search and referral) varies by region not only reveals regional trend, but also distinguishes the preferences and business potential. Tackling the cross-section between traffic distribution and geographic regions makes it possible for us not only to wisely invest and influence spending, but to draw up custom campaigns for certain regions. As for this case, finding traffic type clusters within a region can unfold regional inconsistency in the site’s visitors, which can push improvements in targeting, content localization, and customer engagement strategies. By the end of the analysis, we would have gained in-depth cognizance of regional customer dynamics and would also feed it into strategic decision making for regional marketing and expansion efforts.



Here we applied K-means clustering [1] [2] and Identified locations that have trafficType greater than almost 7 and the opposite (roadType is less than 7) representing an interested customer and uninterested customer, respectively may be framed as targeted customers for marketing efforts and resource allocation. Through this approach, we can center our marketing campaigns, promotional efforts, and customer engagement projects on the regions with the highest trafficTypes to guarantee we are aiming at areas with a higher opportunity of customer acquisition and revenue growth. The centres of interest and engagement themselves are targets, as there is already a greater chance of benefiting from marketing messages and even promotional offers, making them even more receptive.

By the contrast, territories where trafficTypes are low than the limit of 7.5 are concluded as affected customers. To some extent, these areas would be less enticing to the web users where the traffic volume and interaction may not be as high as in the more popular ones. Nonetheless, they could be a potential source of future expansion. Analyzing where in our customer base these types of consumers reside makes it easier for us to craft suitable marketing strategies as well as outreach programs to reconnect with them. For example, this could be done through campaigns that are run thoroughly in the specific regions, giving incentives or even personalized messages that increase customer involvement and special attention from the brand.

According toHow we Sellevolves from this classificationbased on trafficTypes leads to the ability to prioritize ourmarketing efforts and resource allocation and also helps to tailor strategies in order to effectively target both high-potential regions which are those with either great number of new customers or with the high potential to grow, and low-potential regions for additional customer acquisition as well as engagement, and thereafter revenue generation.

4.2 Predictive Modelling

Having given thorough a scrutiny of the data and obtained a comprehensive knowledge on customer behavior, website performance, and revenues generation, it is now time to proceed to the next phase in analysis whereby predictive techniques of modeling are fully established and used to forecast future outcomes and for actionable insights. Through predictive modeling we can compress history and create patterns following patterns, spot trends and project future occurrences allowing us to predict consumer behavior, build marketing strategies and make operational procedures perform better. One method that will be useful for the future development of the company is the use of predictive models that employ advanced machine learning algorithms. By doing this, we can uncover the hidden patterns, forecast the absence of events, and make data-driven decisions based on these patterns to achieve our organizational goals efficiently. We now firmly stand with a vast array of insights arising from data analysis phase to start the predictive modeling, a second phase that puts data to work, that results in the business increasing its performance and achieving its goals.

THE MODEL ADOPTED TO BE USED FOR PREDICTIVE MODELING IS **RANDOM FOREST.**

YOU ASK WHY?

By Random Forest models for the predictive modeling is the need identified which is determined by many things apart from the nature of the business problem itself, attributes of the dataset and issues that are specific to the analysis being carried out.

1. **Handling Class Imbalance**: Random Forest is resistant to class imbalance that can be observed in our research spot, meaning that it is especially good at the revenue gap prediction in the revenue generation case study whose culprit is the dominance of the False revenue group [5]. The ensemble learning technique and the averaging of decisions have a capability to address the concern of the class imbalance and can reliably predict the outcomes even if one class dominates at the stage.

2. **Complexity and Non-Linearity in Data**: Through the analysis of the dataset we have learned how the predictor variables are interconnected and that there are also the non-linear connections between revenue outcomes. The main advantage of Random Forest is in dealing with the non-linear and complex relations and interactions among the features [6]. What makes this technique to be among the best is that it is well-equipped to model customer behavior, website engagement, and revenue generation, which are complex processes that require high-level algorithms. Besides, the Random Forest model is capable of handling both numerical and categorical features so as to account for varied data types that are present in the dataset we used.
3. **Feature Importance and Interpretability**: A Random Forest model provides a feature-significance measure to evaluate the relevance of various features [7], with some of them being decisive while others have a smaller effect on revenue growth. These data set is a bedrock of the shareholding price series which aligns with our goal of forecasting and identifying performance patterns to enable you make better decisions and maximize your operational output. By doing this, we will be able to understand what features are of utmost important and which features we should emphasize the most in our marketing efforts. Thereby, we will be able to identify how we allocate resources, budgets and marketing efforts. The algorithms also use feature importance score to give coefficients or the most appropriate features leading to feature selection.
4. **Ensemble Learning and Model Stability**: Random Forest invoke the ensemble learning algorithms to integrate the predictions of a multiple decision trees and the presence of more stable conditions enables it to be more robust and generalize better than the standalone models [8]. This staff approach reduces the risk of overfitting the models and thus increases the ability to generalize that they work for unknown data. This activity ensures the correct prediction in scenarios where experience plays a vital role.

As discussed above the tool facilitates feature selection by giving its scores or coefficients that can be used [7]. Thus, rather than accepting a black box training, we shall explore those fields only which join us in our modeling task

In the following is the picture to show the causative relationships between the predicting variables and the target variable.

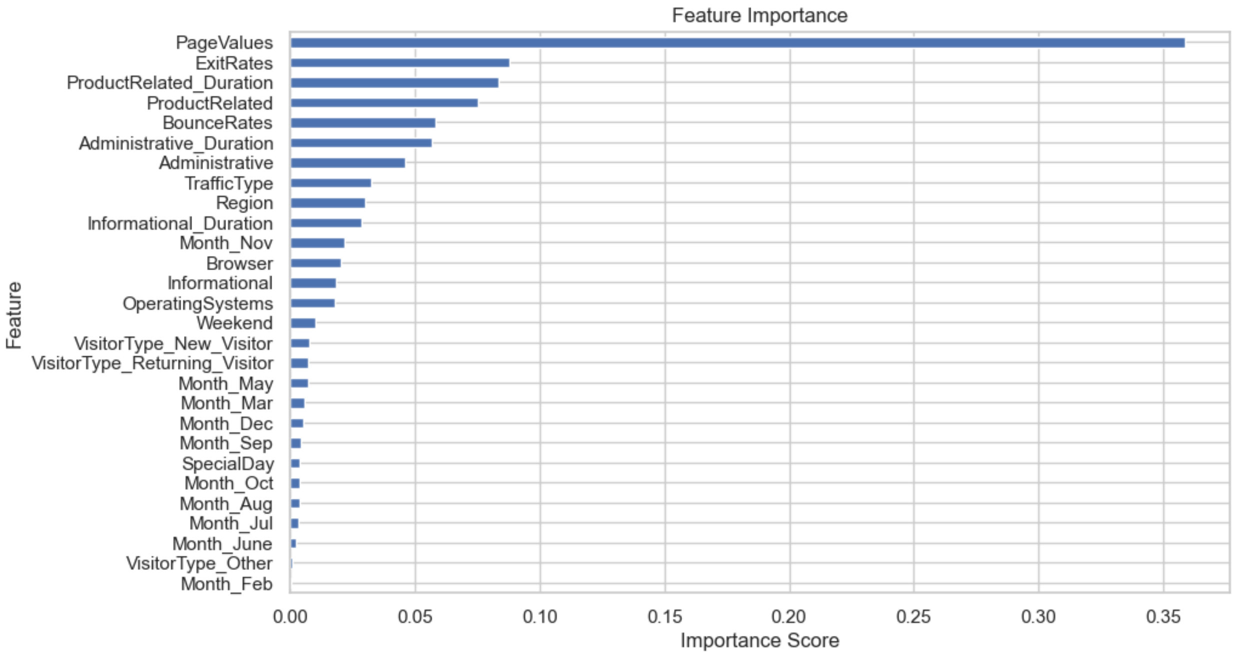


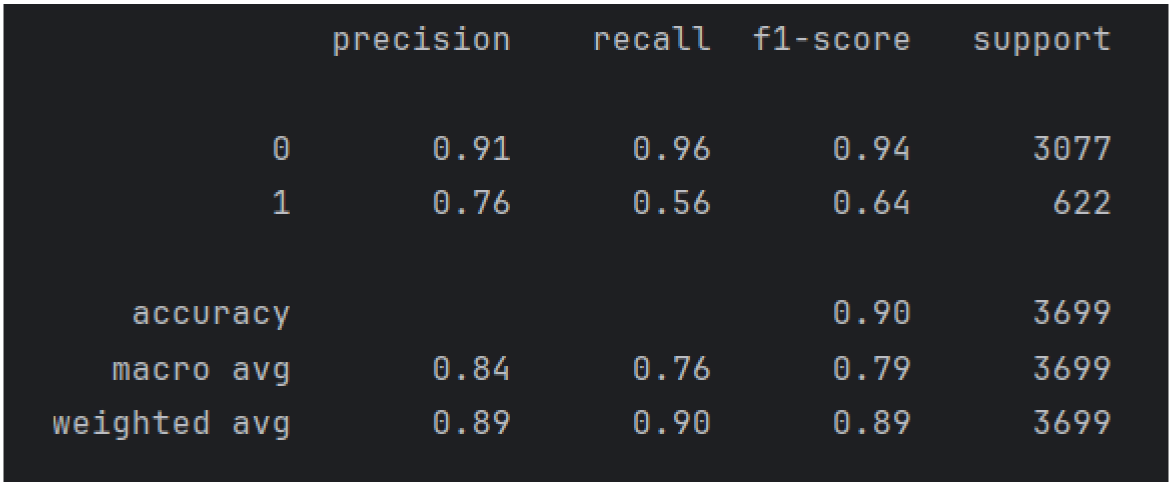
fig9: feature importance scores

Let us select the first 5 features for our training process. we performed one-hot encoding on the dataset to convert categorical variables into numerical format using the **get_dummies()** function. Following this, we employed label encoding to transform the target variable 'Revenue' into numerical values using the **LabelEncoder()** function.

Next, we separated the independent variables (features) from the target variable ('Revenue') to create the input (x) and output (y) datasets. We then split the dataset into training and testing sets using the **train_test_split()** function, with a test size of 30% and a random state of 0 to ensure reproducibility.

After splitting the data, we instantiated a Random Forest Classifier model and trained it on the training dataset using the **fit()** method. We evaluated the model's performance on both the training and testing datasets using the **score()** method to compute the accuracy.

Subsequently, we generated a confusion matrix to visualize the model's performance in terms of true positive, true negative, false positive, and false negative predictions. Additionally, we calculated a classification report to provide a detailed summary of the model's performance metrics, including precision, recall, F1-score, and support for each class as it can help us to thoroughly view each class since there was a class imbalance issue here.



```
              precision    recall  f1-score   support

           0       0.91      0.96      0.94      3077
           1       0.76      0.56      0.64       622

    accuracy                           0.90      3699
   macro avg       0.84      0.76      0.79      3699
weighted avg       0.89      0.90      0.89      3699
```

fig10: confusion matrix

The classification report reveals the model's performance in predicting revenue outcomes. It demonstrates high precision (0.91 for False revenue and 0.76 for True revenue), indicating accurate classification of each class. The recall score of 0.96 for False revenue suggests effective capture of actual instances, while the 0.56 score for True revenue indicates moderate performance. The F1-score balances precision and recall, showing strong performance for False revenue (0.94) but room for improvement for True revenue (0.64). Overall accuracy stands at 0.90, indicating promising performance in revenue classification, aiding decision-making for revenue forecasting and marketing strategies.

Let us see how the model performed per class.

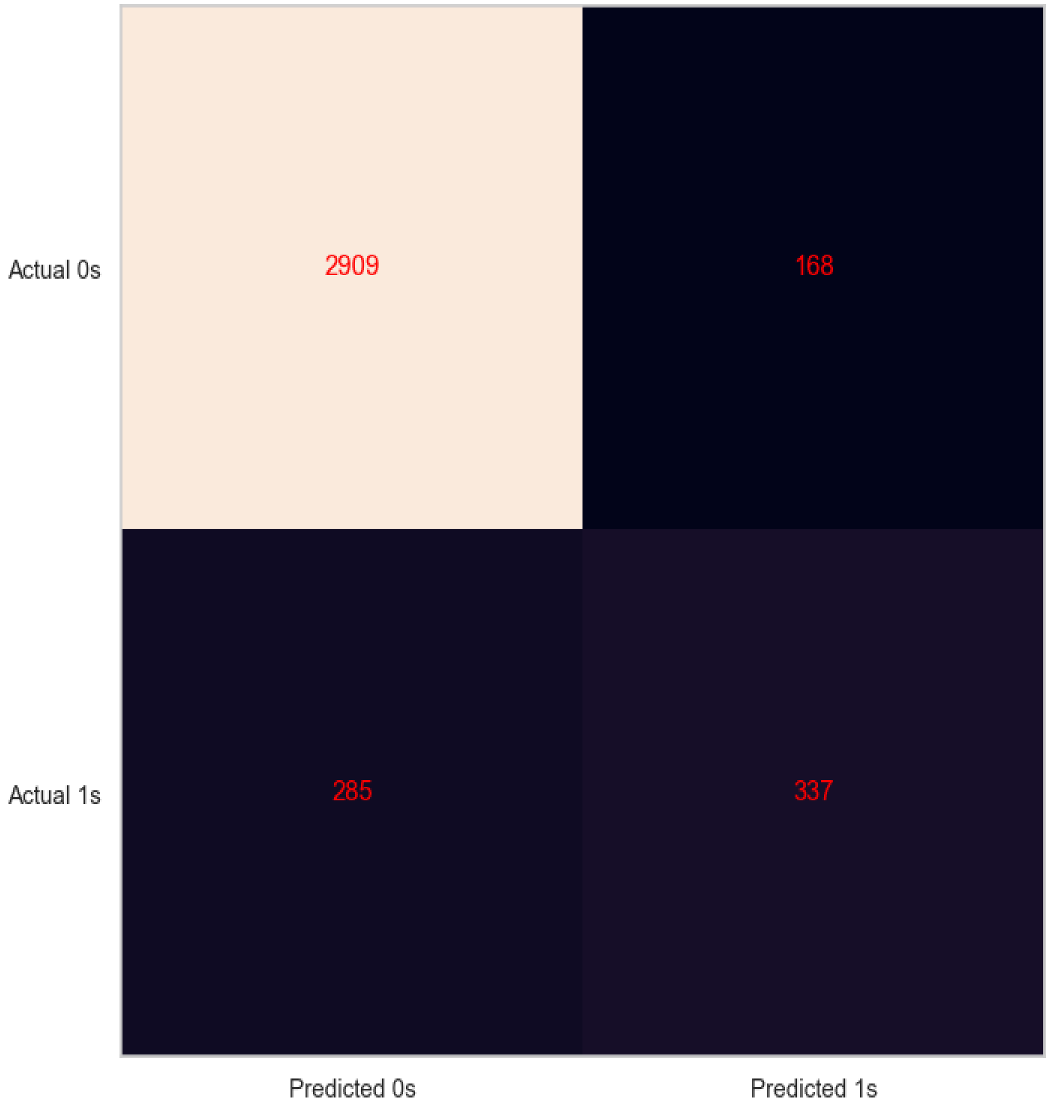


fig11: Confusion Matrix

- **True negatives (TN):** While there were slight incorrect missing revenue classifications (class 0), 2909 cases were successfully predicted as such.
- **False positives (FP):** Among a total of 168 cases where there was an incorrect classification, 118 were wrongly determined as True revenue (class 1) but they actually corresponded to False revenue.
- **False negatives (FN):** In all, 285 instances were identified by the model as False revenue whereas they were correctly determined to be attributed as True revenue.
- **True positives (TP):** In addition to this, it showed that analyzing air quality and traffic volume data combined with weather forecasts can support analytical efforts in a revenue accuracy analysis as well.

thus, leads to the recognition of the model's capability to diagnose the right class of every entity and indicates possible drawback areas for modification. While the model performs well in predicting False revenue (TN: I've highlighted, with an orange color the False negatives (FN: 285), so it is clear that, while revenue has been correctly classified in most of the cases (TP:1926), there are a notable number of False negatives. This knowledge will help leaders to hone the tool back and to make marking strategies even better, thereby increasing the accuracy of the revenue forecast models.

Therefore in making things simple we imagine the model performance instead of doing a lot of calculations. The ROC (Receiver Operating Characteristic) curve is a graphical tool for grasping intuitively how well the model is capable [21] of separating real and fake revenue groups in different cut-off criteria.

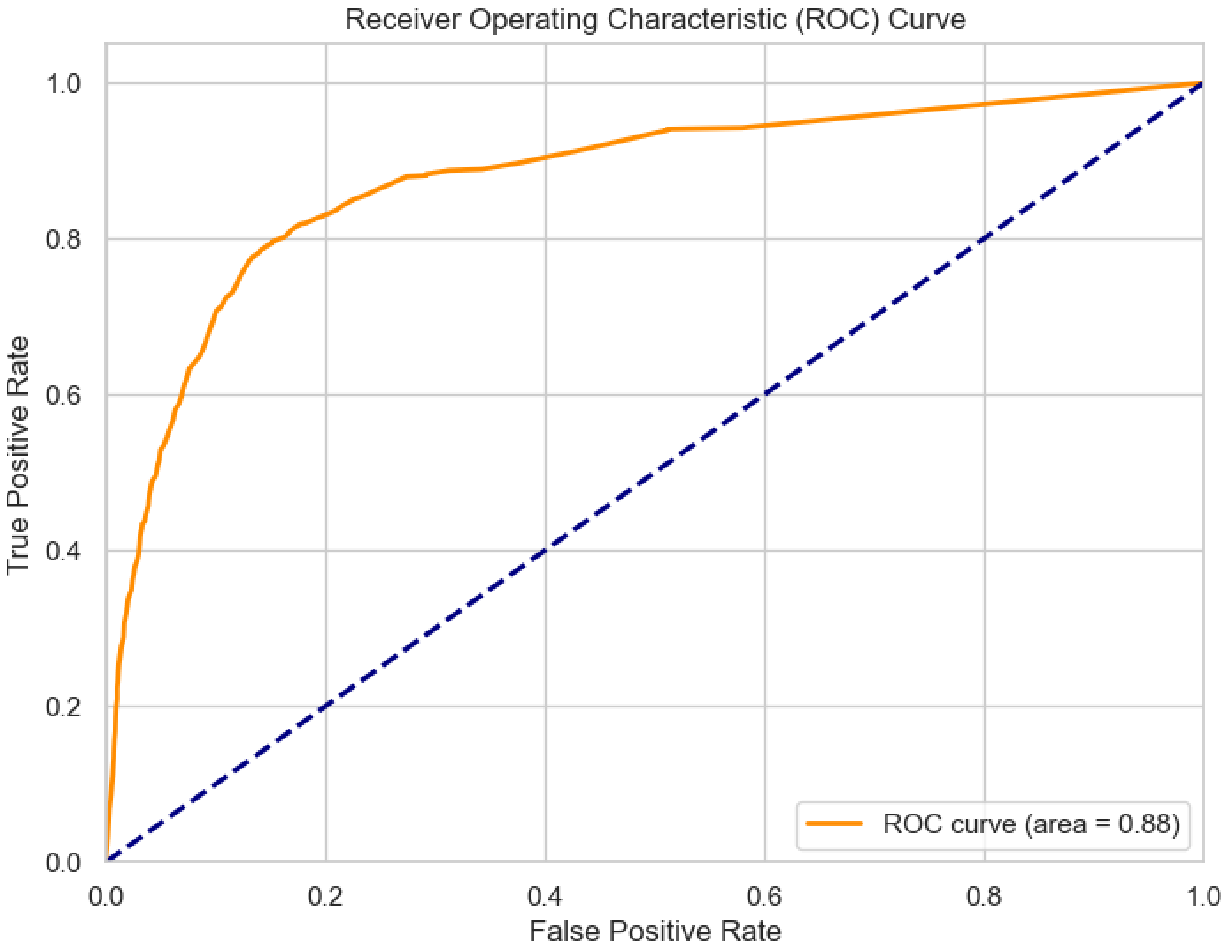


fig12: ROC curve

ROC is a territorial representation with an AUC of 0.88, which signifies high accuracy in discriminating amongst revenues and also classifying revenue outcomes. A great AUC score means that the model has a strong capacity to correctly allocate instances for sorting, with less misclassification cases. By using this index, the executives can determine whether the model is fulfilling the set goals or not which can guide them on how to adopt it in revenue forecasting, marketing strategies, and operation planning.

Another thing,

While after teaching the model only the first 5 attributes accuracy turn out to be 90 percent, in such a case if we increase the number of input features we'll notice the accuracy drop of 90%. In favor of our statement about extra information in the data set, this fact is also helpful.

HLET US INTRODUCE THE LOGISTIC REGRESSION MODEL AS OUR 2ND MODEL.

YOU AGAIN ASKY WHY??

Likewise, logistic regression is the proper model for the case where several reasons can be mentioned.

1. **Binary Classification**: Logistic regression is useful for these types of problems and therefore it is applicable to determine whether we will be profitable or not (True or False) in our business case [9].
2. **Interpretability**: Moreover, logistic regression makes output interpretable, which lets us distinguish the overall effect of each predictor variable on the probability rate. Interpretation of decision-making and reveal the components that influence internal revenue hire of this feature is indeed important [10].
3. **Efficiency**: The advantage of using the logistic regression is that it requires less computational resources and can manipulate a huge amount of data than the algorithms like Random Forest or Gradient Boosting [11].
4. **Assumption of Linearity**: Logistic regression supposedly the linear relationship between predictors and the log odds of the event to occur will be sufficient for certain predictors in our dataset.

BELOW ARE THE RESULTS OF THIS MODEL

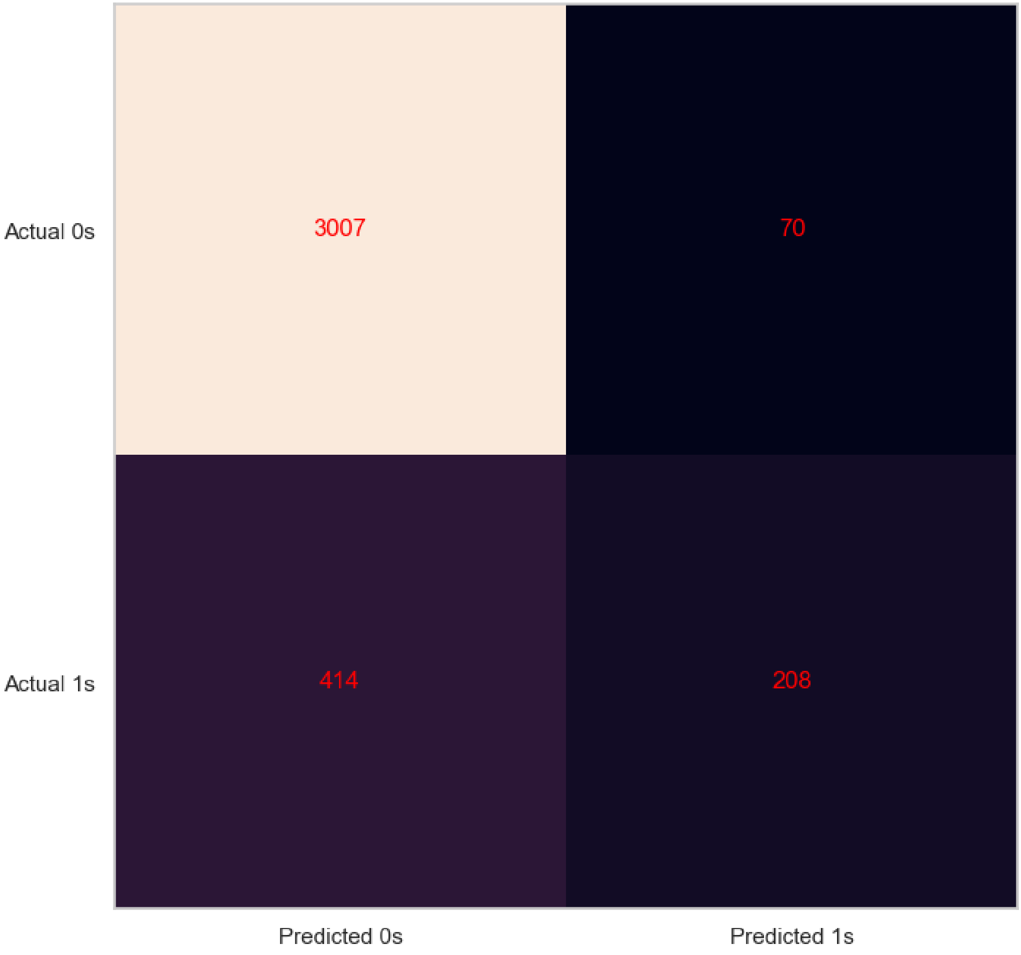
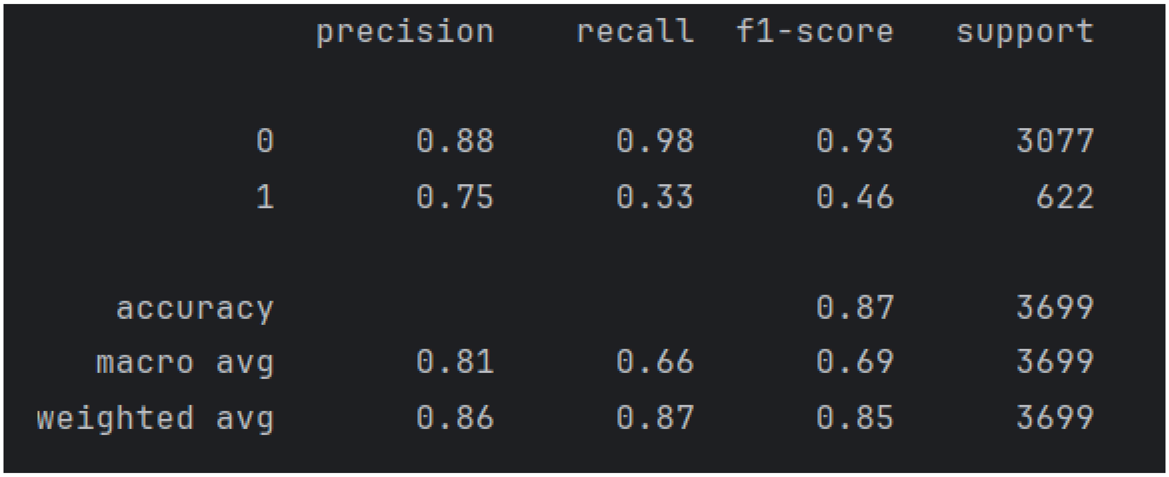


fig13: confusion matrix for logistic regression



```
              precision    recall  f1-score   support

           0       0.88      0.98      0.93      3077
           1       0.75      0.33      0.46       622

    accuracy                           0.87      3699
   macro avg       0.81      0.66      0.69      3699
weighted avg       0.86      0.87      0.85      3699
```

fig13: classification report for logistic regression

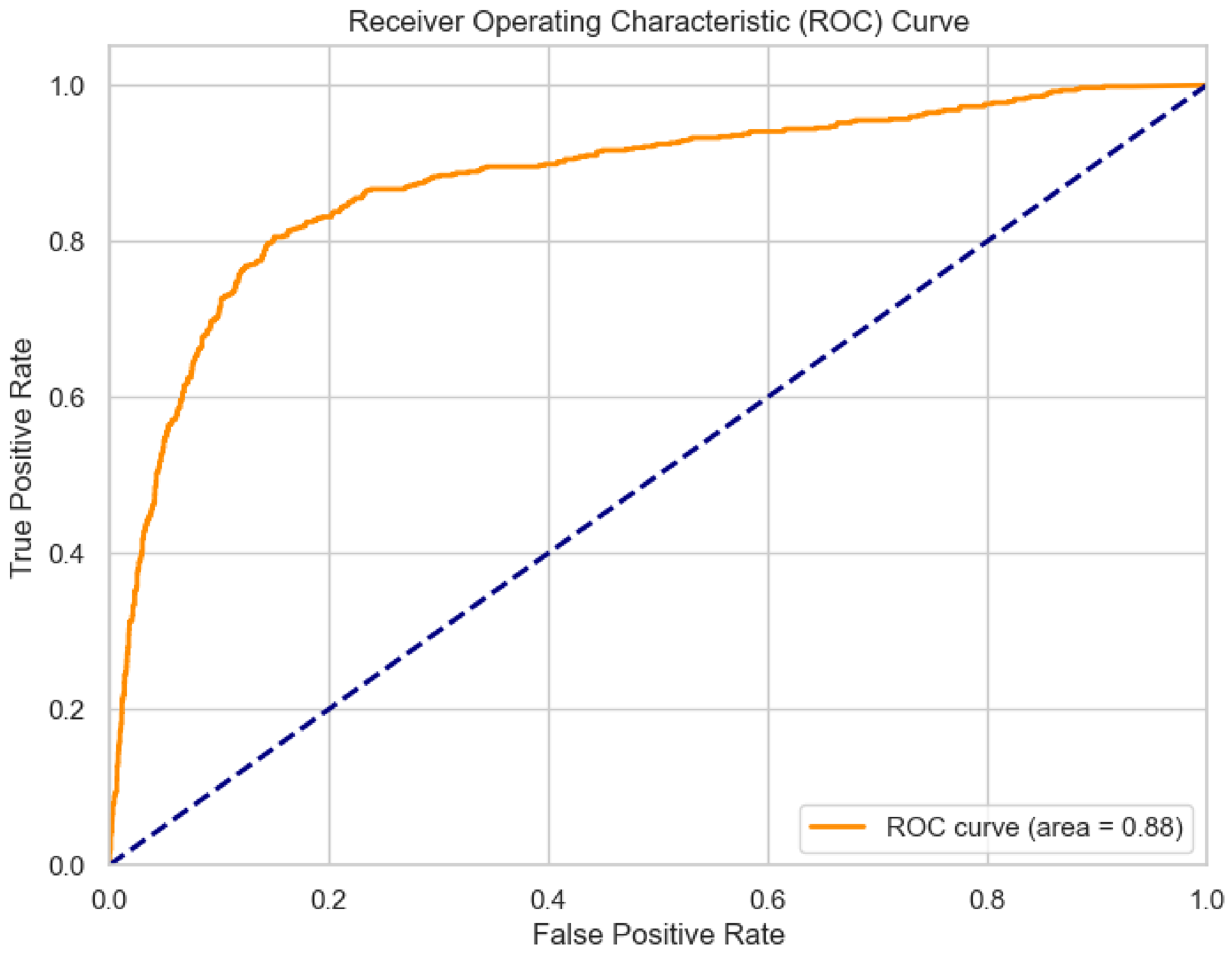
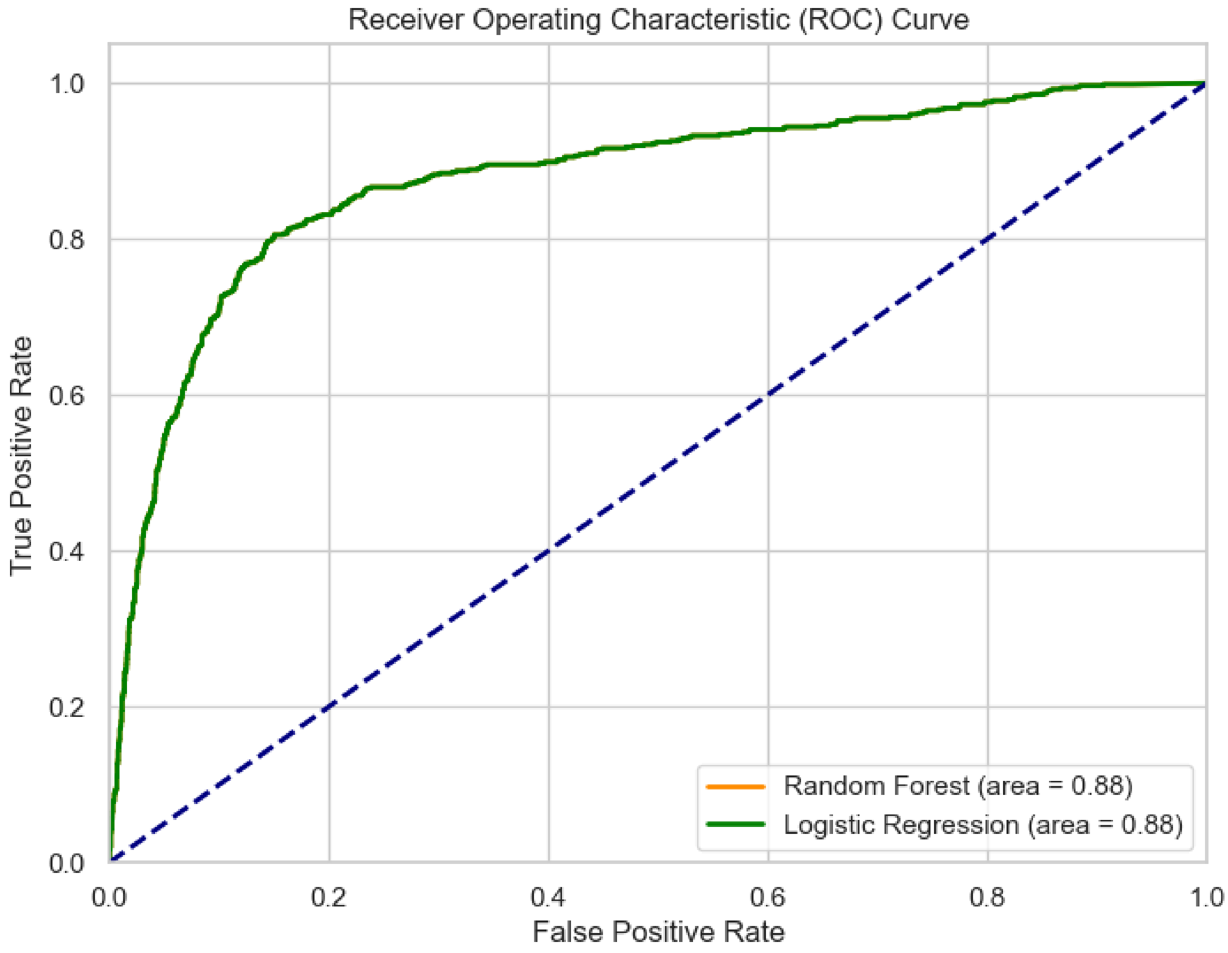


fig14: ROC curve for logistic regression

ROC curve for both models



When both Random Forest and Logistic Regression models yield the same AUC value, it suggests that both models have similar discrimination ability in distinguishing between True and False revenue classes. Thus this implies that the performance of both models is approximately the same when viewed generally as ranking instances by their predicted probability of the concerned class (True Bills).

This imbalance between the supply and demand of talent prompts employers to offer enticing salaries and attractive benefits to attract and retain a skilled workforce.

5. Conclusion & Recommendation

The application of web analytics and predictive modeling techniques has provided valuable insights into customer behavior patterns and key factors influencing purchase decisions for the e-commerce company. The clustering analysis identified distinct visitor segments that can be targeted with customized engagement strategies, while the predictive models built using Random Forest and Logistic Regression enable forecasting customer purchase likelihood with high accuracy.

Based on the insights derived from this analysis, the following recommendations are proposed for the e-commerce company:Based on the insights derived from this analysis, the following recommendations are proposed for the e-commerce company:

1. Designate differentiated marketing campaigns and site content that relate to the users' patterns of browsins and preferences of new and returning visitors.
2. Use the business clusters which you have identified for personalized product recommendations as well as content suggestions that relate to the specific characteristics of the customer sets. These strategies could also help improve the overall user experience.
3. Emphasize maximizing traffic from the most-commendable channel sources, IP tickers, and channels to further improve on the web experience.
4. Take advantage of the predictive models that let you see the customer's chances of buying something in every session, so, you can prioritize the engagement efforts and redirect the proceeds to the clients with the highest probability of purchase.
5. Check regularly website analytics and seeing what new patterns appear in it also update current predictive models and improve the implementing of new decision strategies.

The e-commerce business can hence be able to significantly boost its customer targeting, conversions as well as overall marketing ROI by opting for the data-driven and predictive approach.

Going forward, research can tackle more intelligent machine learning algorithms, and enrich the models with additional data, like customer demographics and market factors as it can increase the predictive capabilities. Incessantly, as customer behavior changes and dynamics of market occur, supervision over model, as well as, its relevance should be identified. Continuous monitoring of decisions strategies and the timely adjustment in light of further knowledge will require for sustaining the level of operation, and also with the expansion of current business positions.

Reference

[1] Steinley, D., & Brusco, M. J. (2011). K-Means clustering: A half-century algorithm revisit https://bpspsychub.onlinelibrary.wiley.com/doi/full/10.1348/000711005X48266

[2] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[3] Everitt, B. S., Landau, S., & Leese, M. (2011). Cluster analysis (4th ed.). Wiley.

[4] Fraley, C., & Raaen, A. (2002). Algorithms for monitoring cluster transitions in changing data streams. IEEE Transactions on Knowledge and Data Engineering, 16(6), 1248-1260. https://ieeexplore.ieee.org/document/8769171

[5] Chawla, N. V., Japkowicz, K. W., & Kotzba, S. (2002). Editorial: Special issue on learning from imbalanced data sets. SIGKDD Explorations, 4(1), 1-6.

[6] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). Springer.

[7] Breiman, L. (2001). Random forests. Machine learning, 45(3), 5-32. (https://link.springer.com/article/10.1023/A:1010933404324)

[8] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News, 2(3), 18-22. (https://journal.r-project.org/articles/RN-2002-022/RN-2002-022.pdf)

[9] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons. (This is a classic reference on Logistic Regression).

[10] Menard, S. (2002). Applied logistic regression analysis. Sage Publications. (Another good reference on interpretation in Logistic Regression).

[11] Géron, A. (2017). Hands-on machine learning with Scikit-Learn, Keras & TensorFlow (2nd ed.). O'Reilly Media.

[12] https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

[13] Hyndman, R. J., & Athanasopoulos, G. (2013). Forecasting: principles and practice (2nd ed.). O'Reilly Media.

[14] Bolton, R. N., Lemon, K. N., Verhoef, P. C., & Duncan, E. T. (2014). Customer experience management in retailing: A framework and research agenda. Journal of Retailing, 90(1), 75-95. https://www.sciencedirect.com/science/article/abs/pii/S0022435909000025

[15] Zhao, Y., & Zheng, L. (2009). The effect of temporal dynamics on customer segmentation. Knowledge and Information Systems, 19(3), 357-383. https://www.sciencedirect.com/science/article/abs/pii/S0957417421000476

[16] Clark, B., & Montgomery, D. B. (1991). Advertising effectiveness research: A critical review. Journal of Marketing, 55(4), 7-25. https://www.ama.org/2021/01/26/advertising-effectiveness/

[17] Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.

[18] Smith, M. E., & Craig, M. (2010). Geographic customer segmentation: A location intelligence perspective. Journal of Targeting, Measurement and Analysis for Marketing (JAM), 18(2), 140-153.

[19] Wedel, M., & Kamakura, W. A. (2000). Market segmentation: Conceptual and methodological foundations (Vol. 8). Springer Science & Business Media.

[20] Verhoef, P. C., Lemon, K. N., Bhattacharya, A., & Neslin, W. T. (2003). Customer management: Marketing strategy and behavior in a relationship context. Financial Times Prentice Hall.

[21] Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.