TURING COLLEGE

CUSTOMERS SEGMENTATION & PURCHASE PROBABILITY PREDICTION

Denis Senin

DA 2023 May

Vilnius, 2024

Table of Contents

Executive summary

The analytics data pertains to a fictional company specializing in online retail of electronics. The data source for this project is a publicly available dataset obtained from Kaggle.com, accessible here.

The primary objective of the company is to enhance customer experience, elevate satisfaction levels, and ultimately boost revenue. To achieve these goals, the project employed various data analysis techniques, including RFM segmentation, assessment of customer lifetime value (CLV), and analysis of purchase probability.

The project's intended audience is the sales and marketing executives of the company, keen on comprehending the company's sales performance and pinpointing areas for improvement. It aims to deliver detailed insights into various facets of the company's operations, such as sales and marketing trends, customer analysis and performance improvement.

Ultimately, the goal is to provide actionable insights to the marketing and sales executives, facilitating improved company operations and fostering growth.

The project aimed to answer several questions, including:

- Who are the company's most valuable customers, and how can the company segment them to improve the effectiveness of targeted marketing campaigns?
- What is the Customer Lifetime Value (CLV), and how does it vary across different countries? How can the company enhance its value?
- How can the company estimate the customers' probability of purchase based on their behavioural characteristics, and what marketing strategies can be used to improve targeting of such customers?

Introduction to the dataset

Company Google analytics dataset which provides the customers' information about their behaviour on the website and acquisition sources. It includes over 243,297 records and 55 attributes for the period from 2016-08-01 to 2017-08-01. Dataset is designed for data scientists and divided into test and training datasets.

Considering what dataset to choose the main criteria was the one with more variables available (55 in the train dataset vs 53 in the test dataset). Though still both datasets are quite extensive and have a large size that make them impractical for such platforms like Big Query (upload limitation of 100 MB). Analysing a smaller timeframe enhances manageability, computational efficiency, and resource optimization while still maintaining representativeness through careful sampling. This approach enables focused analysis to derive meaningful insights despite the constraints of working with a reduced dataset. Therefore, a three-month subset (from 2016-08-01 to 2016-10-31) was chosen for further analysis. Throughout this period, the company generated a revenue of approximately $394,03 B, engaging 199,6 K unique visitors in roughly 2536 distinct transactions, with an average revenue value per user of approximately $138,26 M.

Data preparation and preprocessing

Before commencing any data analysis, data preparation and preprocessing steps were executed to guarantee the accuracy and consistency of the dataset. This process entailed identifying and rectifying duplicate records, as well as ensuring the appropriate data types were utilized and additional fields were created where necessary.

Main analysis:

Feature Engineering - 'has_transaction' Field Creation: To enhance analytical capabilities, a new field named 'has_transaction' was introduced. This binary indicator (1 for records associated with revenue-generating transactions and 0 otherwise) was populated based on the presence of non-null values in the 'totals_transactionRevenue' field. This addition significantly streamlined subsequent analyses, aiding in the segmentation of transactional data and enabling focused investigations into user behaviour and revenue patterns.

Probability of purchase analysis:

Handling Missing Values and Duplicates: Checks were performed to identify and address missing data and duplicates. Missing values were either imputed (e.g., using median or mode) or removed as appropriate to maintain data integrity.

Outlier Identification and Management: Outliers were identified and managed to normalize the distribution of variables, ensuring the robustness of subsequent analyses.

Variable Encoding: Categorical and Boolean variables were appropriately encoded to prepare them for logistic regression modelling, facilitating the inclusion of these features in predictive analyses.

Feature Selection Based on Correlation and Significance: Variables were carefully selected for modelling based on their correlation with the target variable (probability of purchase) and considerations of multicollinearity. This step ensured that the final model included the most relevant predictors for accurate and interpretable results.

Analysis

The project commences with descriptive dataset analysis to identify patterns and dependencies between purchase behaviour and various variables such as country, marketing channel, and technology preferences. Subsequently, RFM segmentation is performed to categorize customers based on their recency, frequency, and monetary value. CLV calculations are then conducted to estimate the lifetime value of customers. Finally, predictive models employing logistic regression is developed to predict the probability of purchase.

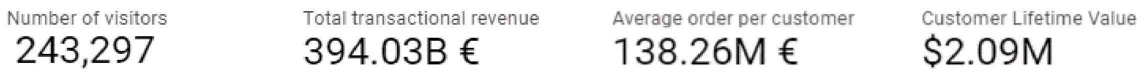
The analysis encompasses several stages:

- Descriptive Dataset Analysis: Explores dependencies between purchase behaviour and various customer attributes.
- RFM Segmentation: Categorizes customers into segments based on recency, frequency, and monetary value.
- CLV Calculation: Estimates the lifetime value of customers to identify high-value segments.
- Predictive Modelling: Utilizes logistic regression to predict the probability of purchase.

1. Descriptive dataset analysis

Once the data was prepared and pre-processed using SQL, descriptive dataset analysis was conducted to provide an overview of the company's performance and gain initial insights. The visualizations were prepared using the business intelligence tool Looker Studio, where interactive dashboards were created. You can access the interactive dashboard through this link.

From these data points, we observe that we have 243,297 unique customers, generating approximately $3.94 B in revenue from 2016 August to 2016 October, with an average order value of around $138,26 M. Utilizing the predictive model, it was determined that our predictive customer lifetime value is $2,6 M for the specific time.



Number of visitors	Total transactional revenue	Average order per customer	Customer Lifetime Value
243,297	394.03B €	138.26M €	$2.09M

Chart 1.1 Main KPI’s

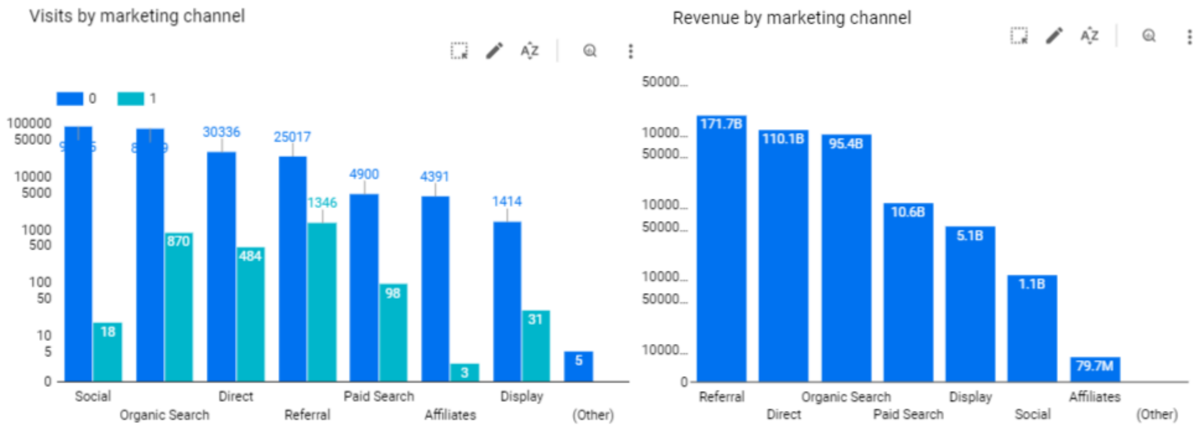


Chart 1.2. Visits and revenue by the marketing channel

Referral and Organic Search have relatively higher transaction rates compared to other channels. Paid Search, Display, and Affiliates have low transaction rates. As for social media channels - majority of visitors from social channels do not convert into transactions.

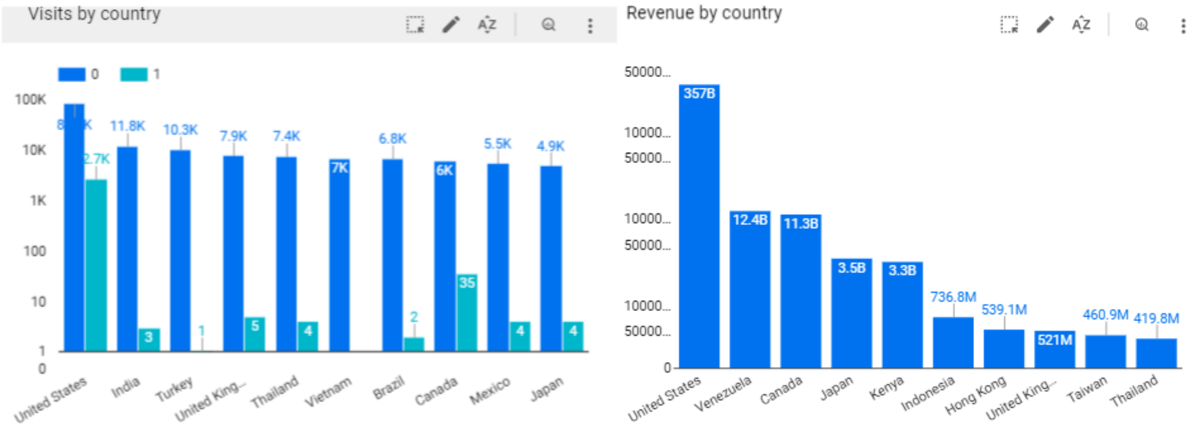


Chart 1.3. Visits and revenue by country

The United States, India, and other countries like Turkey, United Kingdom, and Canada had a high number of visitors, but a majority did not convert into transactions. The conversion rates vary significantly across countries. For instance, the United States had a relatively higher conversion rate compared to India, where most of visitors did not transact. The United States leads in transactional revenue with a significant amount of $357 B indicating a substantial contribution to overall revenue. The United States contributes the highest transactional revenue, indicating a strong conversion rate despite a relatively lower number of transacted visitors compared to total visitors.

Canada shows a relatively high transactional revenue per visitor, suggesting a higher conversion rate compared to other countries with similar visitor counts. Despite many visitors, India has a low conversion rate, indicating potential challenges in converting visitors into customers. Canada and Japan, show relatively high revenue per visitor and such countries like Mexico, Thailand show moderate transactional revenue and visitor counts.

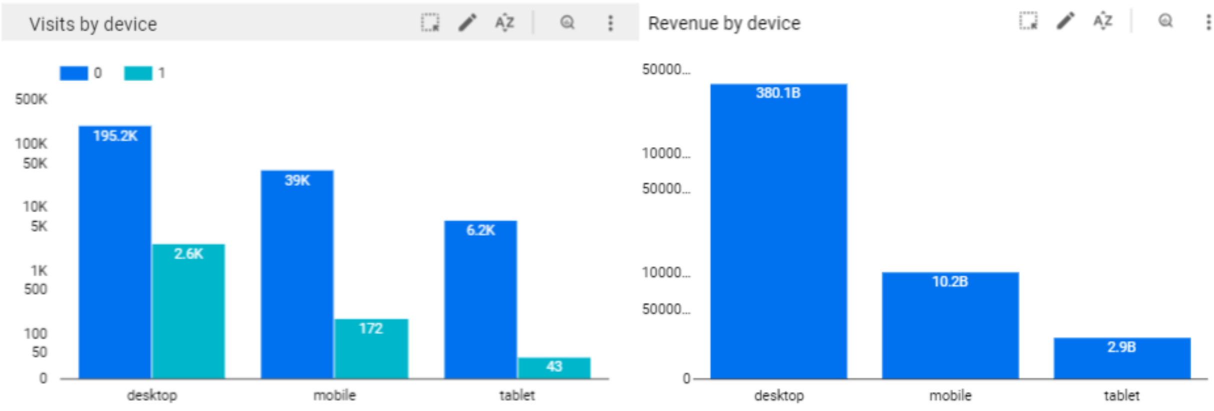


Chart 1.4. Visits and revenue by device

Desktop users account for most visitors, suggesting that desktop remains a primary platform for accessing the website or platform. The conversion rate (transacted visitors/total visitors) for desktop users is relatively lower compared to the total number of desktop visitors.

Mobile users represent a significant audience segment but the conversion rate for mobile users is relatively low compared to the total number of mobile visitors.

Tablet users constitute a smaller segment of visitors compared to desktop and mobile. The conversion rate for tablet users is higher compared to mobile users but still relatively low compared to desktop users.

Desktop users contribute the highest transactional revenue, mobile users also generate a significant amount of transactional revenue but both channels have a relatively lower conversion rate (transacted visitors/total visitors). Tablet users contribute a smaller portion of both visitors and transactional revenue compared to desktop and mobile users but the conversion rate among tablet users is relatively higher compared to mobile users.

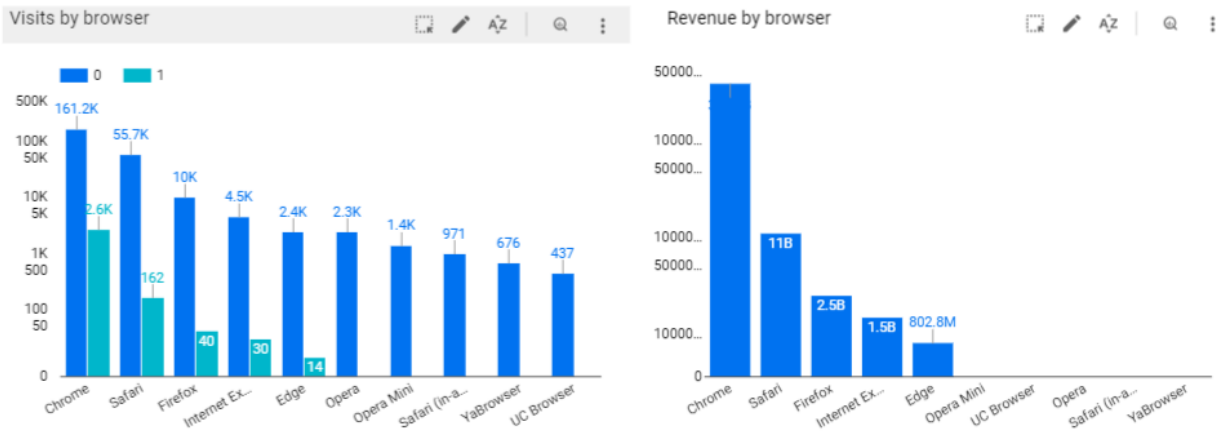


Chart 1.5. Visits and revenue by browser

Analysing split by browsers: Chrome contributes the highest transactional revenue among all browsers, reflecting its popularity and widespread usage among visitors. Safari contributes a significant amount of transactional revenue, although lower than Chrome, indicating strong purchasing power among Safari users. Firefox and Internet Explorer generate moderate transactional revenue compared to Chrome and Safari, suggesting a smaller contribution to overall revenue, and showing lower usage and conversion rates compared to more popular browsers. Edge generates a relatively lower amount of transactional revenue compared to other browsers, indicating a smaller audience and potentially lower conversion rates.

For further analysis f.e. for profiling of the valuable customer, it's important to consider a range of qualities and attributes that can help you identify and target your most valuable and receptive audience:

- Demographic information (age, gender, income level, education)
- Psychographic characteristics (lifestyle, values and beliefs, personality traits)
- Behavioural patterns (purchase and online behaviour, brand interactions)
- Geographic factors (location, urban vs rural)
- Technographic details (device preferences, tech adoption).
- Customer pain points and needs (challenges, needs and desires)
- Purchase intent and decision-making process (buying motivations and decision-making factors)
- Customer Lifetime Value (CLV) (revenue potential)
- Customer satisfaction and loyalty (Gauge customer satisfaction and loyalty metrics)

The descriptive dataset analysis shows that for defining valuable customer we might be focused on the customers:

- from countries with the highest transactional revenue and conversion rates, such as the United States and Canada,
- using the most popular browsers like Chrome and Safari, which show significant transactional revenue and visitor volume,
- desktop users, as they contribute the highest transactional revenue, while also optimizing for mobile users given their substantial presence
- customers using such marketing channels like organic search and paid advertising.

Additionally, we will segment customers by their visits or purchase recency, frequency and monetary values. Customer retention

2. RFM segmentation

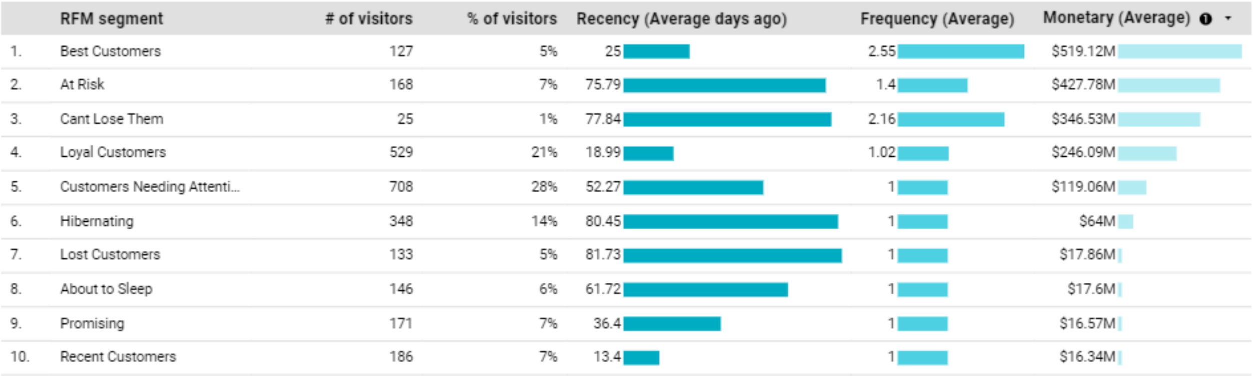
RFM analysis segments customers based on their transactional behaviour, where RFM refers to Recency (time elapsed since a customer's last interaction), Frequency (number of interactions within a specific period), and Monetary (total spending within the same period). This segmentation technique aids in identifying valuable customers for retention and upselling opportunities while also enabling strategies to retain other customers.

The process of assigning scores involved calculating RFM scores for each customer using SQL code. This was done by determining percentiles for the customers' Recency, Frequency, and Monetary values. Customers were then categorized into one of nine segments based on these RFM scores. The code calculated percentiles for Monetary, Frequency, and Recency values by dividing them into five equal parts (25th, 50th, 75th, and 100th percentiles). Subsequently, RFM scores were computed for each customer by comparing their Monetary, Frequency, and Recency values to these percentiles. Finally, customers were assigned to one of the nine segments based on their combined R (Recency), F (Frequency), and M (Monetary) scores.

The average number of pageviews (34) for transacted customers compared to non-transacted customers (4) serves as an indicator of engagement. Further analysis for non-transacted customers can delve into the nuances of this engagement—such as which pages are most viewed, session duration, or frequency of returning visitors—which may reveal insights into potential barriers to conversion.

RFM for Transacted Customers:

For transacted customers, RFM analysis with monetary (revenue) was taken as the primary metric can reveal insights into the most valuable customers based on their purchase recency, frequency of purchases, and total monetary value. This helps identify high-value segments for targeted marketing strategies or retention efforts.



	RFM segment	# of visitors	% of visitors	Recency (Average days ago)	Frequency (Average)	Monetary (Average)
1.	Best Customers	127	5%	25	2.55	$519.12M
2.	At Risk	168	7%	75.79	1.4	$427.78M
3.	Cant Lose Them	25	1%	77.84	2.16	$346.53M
4.	Loyal Customers	529	21%	18.99	1.02	$246.09M
5.	Customers Needing Attenti...	708	28%	52.27	1	$119.06M
6.	Hibernating	348	14%	80.45	1	$64M
7.	Lost Customers	133	5%	81.73	1	$17.86M
8.	About to Sleep	146	6%	61.72	1	$17.6M
9.	Promising	171	7%	36.4	1	$16.57M
10.	Recent Customers	186	7%	13.4	1	$16.34M

Chart 2.1. RFM analysis for transacted customers

Best Customers:

- Recency: These customers were active relatively recently, with an average recency of 25 days.
- Frequency: They transact moderately frequently, averaging about 2.55 transactions per visitor.
- Monetary: They have the highest average monetary value per transaction among all segments, with an average of $519,124,173.20.

Insight: Continue prioritizing personalized engagement and exclusive offerings to maintain and increase their spending.

At Risk:

- Recency: These customers have a higher average recency of about 76 days.

- Frequency: Their transaction frequency is relatively low, averaging around 1.40 transactions per visitor.
- Monetary: Despite lower frequency, they still contribute significantly to revenue, averaging $427,779,404.80.

Insight: Implement re-engagement campaigns and personalized incentives to prevent further disengagement and encourage repeat transactions.

Can't Lose Them:

- Recency: They are relatively inactive with a higher recency of 78 days.
- Frequency: These customers transact more frequently compared to others, averaging 2.16 transactions per visitor.
- Monetary: They contribute a substantial amount per transaction, averaging $346,531,600.00.

Insight: Develop strategies to retain these customers through targeted communications and special offers to maintain their value.

Loyal Customers:

- Recency: Recently active with an average recency of around 19 days.
- Frequency: Their transaction frequency is low, averaging about 1.02 transactions per visitor.
- Monetary: Despite lower frequency, they have a considerable average monetary value per transaction, averaging $246,088,279.80.

Insight: Encourage more frequent transactions through loyalty rewards and personalized promotions to increase their lifetime value.

Customers Needing Attention:

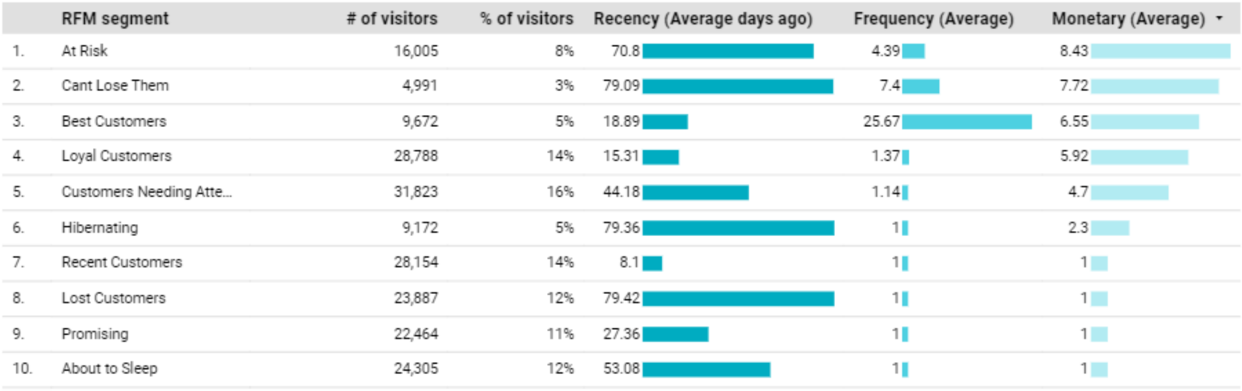
- Recency: Moderate recency of around 52 days.
- Frequency: Like loyal customers, they transact infrequently, averaging about 1.00 transaction per visitor.

- Monetary: Their average monetary value per transaction is moderate, averaging $119,062,909.60.

Insight: Focus on improving engagement and repeat business through targeted marketing and enhanced customer experiences.

RFM for Non-Transacted Customers:

For non-transacted customers, using total pageviews as a proxy for engagement, RFM analysis can highlight patterns such as how recent their engagement was, how frequently they visit company site, and for monetary - the depth of their engagement (e.g., number of unique pages visited). This can guide efforts to convert these engaged visitors into customers.



	RFM segment	# of visitors	% of visitors	Recency (Average days ago)	Frequency (Average)	Monetary (Average) ▾
1.	At Risk	16,005	8%	70.8	4.39	8.43
2.	Cant Lose Them	4,991	3%	79.09	7.4	7.72
3.	Best Customers	9,672	5%	18.89	25.67	6.55
4.	Loyal Customers	28,788	14%	15.31	1.37	5.92
5.	Customers Needing Atte...	31,823	16%	44.18	1.14	4.7
6.	Hibernating	9,172	5%	79.36	1	2.3
7.	Recent Customers	28,154	14%	8.1	1	1
8.	Lost Customers	23,887	12%	79.42	1	1
9.	Promising	22,464	11%	27.36	1	1
10.	About to Sleep	24,305	12%	53.08	1	1

Chart 2.2. RFM analysis for non-transacted customers

At Risk:

- Recency: The average time since their last visit or transaction is relatively high (70.8 days), indicating they may be losing interest.
- Frequency: They visit or transact about 4 times on average.
- Monetary: Moderate number of pageviews.

Insight: Target these customers with re-engagement campaigns or special offers to bring them back to active status.

Can't Lose Them:

- Recency: High average recency (79.1 days), suggesting they are loyal but might need continued attention.

- Frequency: More frequent visitors/transactors.
- Monetary: Higher pageviews numbers compared to some segments.

Insight: Maintain engagement with personalized communications and exclusive offerings to retain their loyalty.

Best Customers:

- Recency: Recently active with an average of 18.9 days.
- Frequency: Very frequent transactions.
- Monetary: Highest number of pageviews.

Insight: Focus on providing VIP treatment, exclusive benefits, and rewards to maximize their value.

Loyal Customers:

- Recency: Recently active like the Best Customers.
- Frequency: Less frequent but consistent transactions.
- Monetary: Moderate engagement.

Insight: Nurture this group with loyalty programs and targeted promotions to encourage more frequent transactions.

Customers Needing Attention:

- Recency: Moderate recency but could be more engaged.
- Frequency: Lower frequency of transactions.
- Monetary: Moderate spending.

Insight: Implement strategies to increase engagement and conversion through personalized offers.

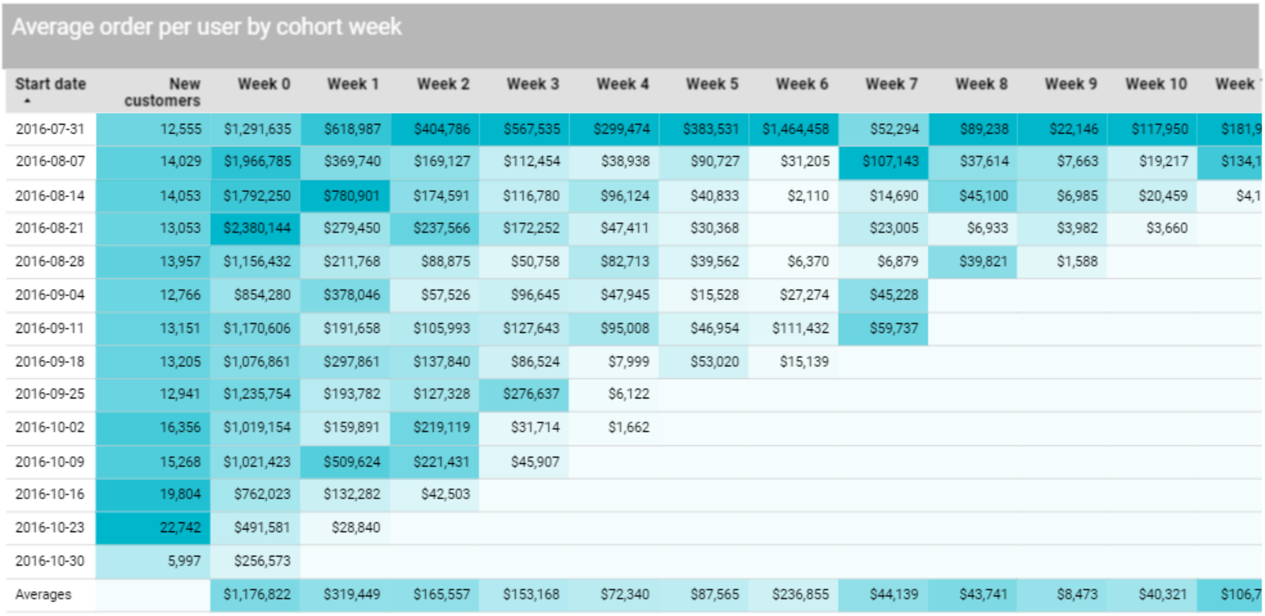
For identifying valuable customers for engagement (and therefore possibly a revenue) maximization, company should target segments with high monetary value and potential for increased frequency and retention. The **Best Customers** and **Can't Lose Them** segments are strong candidates due to their high engagement and relatively frequent visits.

While performing RFM analysis for transacted and non-transacted visitors the categories we can consider as valuable customers are: **Best Customers**, **Can't Lose them, At Risk** and **Loyal Customers**.

3. Customer Lifetime Value Analysis

Customer Lifetime Value (CLV) is a metric used to estimate the total value a customer will generate for the company throughout their lifetime. It helps identify the most valuable customers for retention and upsell opportunities. CLV can be calculated using various methods, including standard formula, historical averages, or predictive modelling. For this analysis, a more in-depth approach such as predictive modelling using historical averages on cohort analysis was utilized, allowing for continuous updates based on new data.

The Customer Lifetime Value (CLV) analysis starts by pinpointing the customer's initial purchase time, marking their introduction to the company for the current year as our starting cohort. This is followed by reviewing all transactions made in the following weeks and calculating the revenue generated during those periods. This revenue is then divided by the cohort size to determine the average order value and assess how it varies over time.



Average order per user by cohort week

Start date	New customers	Week 0	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week
2016-07-31	12,555	$1,291,635	$618,987	$404,786	$567,535	$299,474	$383,531	$1,464,458	$52,294	$89,238	$22,146	$117,950	$181,9
2016-08-07	14,029	$1,966,785	$369,740	$169,127	$112,454	$38,938	$90,727	$31,205	$107,143	$37,614	$7,663	$19,217	$134,1
2016-08-14	14,053	$1,792,250	$780,901	$174,591	$116,780	$96,124	$40,833	$2,110	$14,690	$45,100	$6,985	$20,459	$4,1
2016-08-21	13,053	$2,380,144	$279,450	$237,566	$172,252	$47,411	$30,368		$23,005	$6,933	$3,982	$3,660	
2016-08-28	13,957	$1,156,432	$211,768	$88,875	$50,758	$82,713	$39,562	$6,370	$6,879	$39,821	$1,588		
2016-09-04	12,766	$854,280	$378,046	$57,526	$96,645	$47,945	$15,528	$27,274	$45,228				
2016-09-11	13,151	$1,170,606	$191,658	$105,993	$127,643	$95,008	$46,954	$111,432	$59,737				
2016-09-18	13,205	$1,076,861	$297,861	$137,840	$86,524	$7,999	$53,020	$15,139					
2016-09-25	12,941	$1,235,754	$193,782	$127,328	$276,637	$6,122							
2016-10-02	16,356	$1,019,154	$159,891	$219,119	$31,714	$1,662							
2016-10-09	15,268	$1,021,423	$509,624	$221,431	$45,907								
2016-10-16	19,804	$762,023	$132,282	$42,503									
2016-10-23	22,742	$491,581	$28,840										
2016-10-30	5,997	$256,573											
Averages		$1,176,822	$319,449	$165,557	$153,168	$72,340	$87,565	$236,855	$44,139	$43,741	$8,473	$40,321	$106,7

Chart 3.1. Customers average order value by cohort size

Table 3.1 illustrates the average revenue per customer for various weekly cohorts from August 2016 to October 2016. Each row corresponds to a cohort's initial purchases, and each column indicates the average weekly revenue for that cohort relative to the starting week. For instance, in the beginning of August 2016 (Week 0), the average revenue per customer was ~$1,18 M, while in the second week of August (Week 1), it decreased to ~$618 K. The table demonstrates the changes in average revenue over time.

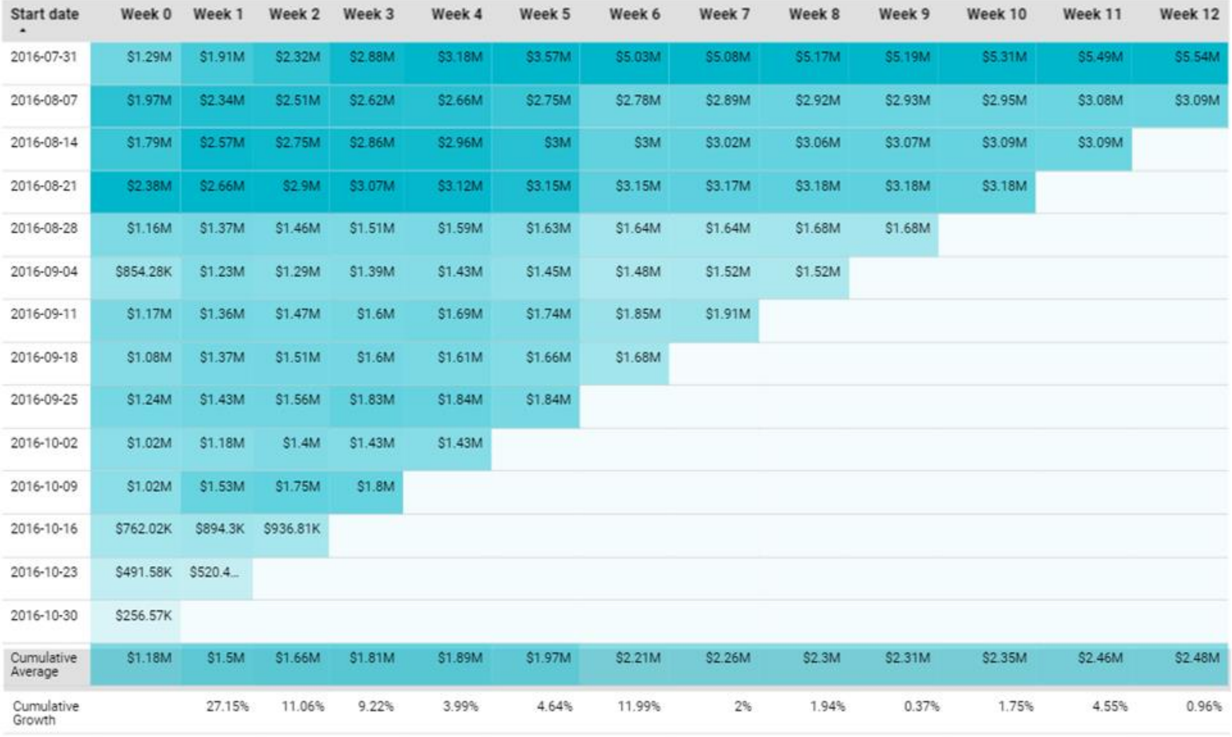
Fluctuations in Revenue: There is a significant variance in revenue from week to week across different start dates (cohorts). For example, in Week 0, the revenue ranges from a high of approximately $2.38 M (2016-08-21) to a low of around $256,573 (2016-10-30).

Initial Revenue Impact: The revenue in Week 0 (the first week after registration) tends to be notably higher compared to subsequent weeks. This suggests that customers may make larger purchases or more frequent purchases shortly after registering.

Decline Over Time: Like the CLV trend, there is a general decline in revenue over the subsequent weeks following registration. This decline could be indicative of reduced customer engagement or spending as time progresses post-registration.

Seasonal Patterns: There are instances of lower revenue during specific periods, possibly reflecting seasonal influences or external factors affecting customer behaviour and purchasing patterns.

Average Revenue: The average revenue by week steadily decreases from Week 0 to Week 12, with the highest average revenue occurring in Week 0 (around $1.18 M) and the lowest in later weeks (e.g., around $23,629 in Week 12).

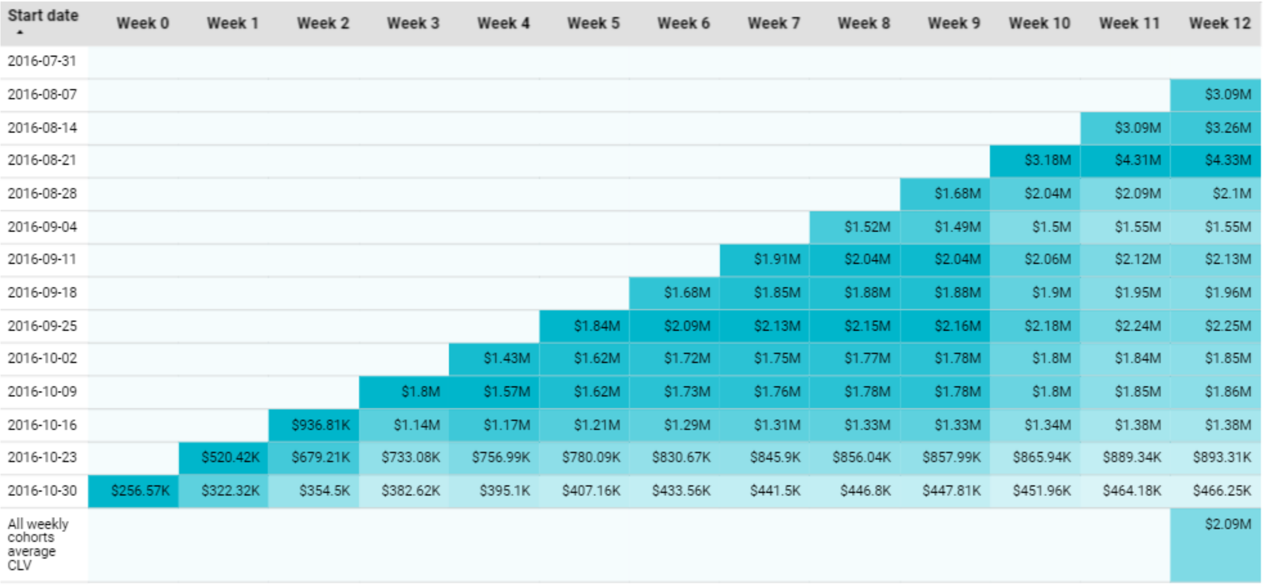


Start date ▴	Week 0	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12
2016-07-31	$1.29M	$1.91M	$2.32M	$2.88M	$3.18M	$3.57M	$5.03M	$5.08M	$5.17M	$5.19M	$5.31M	$5.49M	$5.54M
2016-08-07	$1.97M	$2.34M	$2.51M	$2.62M	$2.66M	$2.75M	$2.78M	$2.89M	$2.92M	$2.93M	$2.95M	$3.08M	$3.09M
2016-08-14	$1.79M	$2.57M	$2.75M	$2.86M	$2.96M	$3M	$3M	$3.02M	$3.06M	$3.07M	$3.09M	$3.09M	
2016-08-21	$2.38M	$2.66M	$2.9M	$3.07M	$3.12M	$3.15M	$3.15M	$3.17M	$3.18M	$3.18M	$3.18M		
2016-08-28	$1.16M	$1.37M	$1.46M	$1.51M	$1.59M	$1.63M	$1.64M	$1.64M	$1.68M	$1.68M			
2016-09-04	$854.28K	$1.23M	$1.29M	$1.39M	$1.43M	$1.45M	$1.48M	$1.52M	$1.52M				
2016-09-11	$1.17M	$1.36M	$1.47M	$1.6M	$1.69M	$1.74M	$1.85M	$1.91M					
2016-09-18	$1.08M	$1.37M	$1.51M	$1.6M	$1.61M	$1.66M	$1.68M						
2016-09-25	$1.24M	$1.43M	$1.56M	$1.83M	$1.84M	$1.84M							
2016-10-02	$1.02M	$1.18M	$1.4M	$1.43M	$1.43M								
2016-10-09	$1.02M	$1.53M	$1.75M	$1.8M									
2016-10-16	$762.02K	$894.3K	$936.81K										
2016-10-23	$491.58K	$520.4...											
2016-10-30	$256.57K												
Cumulative Average	$1.18M	$1.5M	$1.66M	$1.81M	$1.89M	$1.97M	$2.21M	$2.26M	$2.3M	$2.31M	$2.35M	$2.46M	$2.48M
Cumulative Growth		27.15%	11.06%	9.22%	3.99%	4.64%	11.99%	2%	1.94%	0.37%	1.75%	4.55%	0.96%

Chart 3.2. Cumulative average order value by cohort size

Table 3.2 presents the same data as Table 3.1, but the revenue for each cohort in a particular week is expressed as a cumulative sum. This means that the revenue from the previous week is added to the revenue from the current week. Subsequently, the averages for all week numbers (representing weeks since the first purchase) were calculated, followed by the calculation of percentage growth based on these average numbers.

Essentially, the table below displays the revenue growth by customers' first purchase cohort over X weeks after registration. The figures below summarize these values in terms of cumulative revenue averages and percentages (Growth %). This offers a comprehensive perspective on the anticipated revenue growth for the business based on historical data.



Start date ▲	Week 0	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12
2016-07-31													
2016-08-07													$3.09M
2016-08-14												$3.09M	$3.26M
2016-08-21											$3.18M	$4.31M	$4.33M
2016-08-28										$1.68M	$2.04M	$2.09M	$2.1M
2016-09-04									$1.52M	$1.49M	$1.5M	$1.55M	$1.55M
2016-09-11								$1.91M	$2.04M	$2.04M	$2.06M	$2.12M	$2.13M
2016-09-18							$1.68M	$1.85M	$1.88M	$1.88M	$1.9M	$1.95M	$1.96M
2016-09-25						$1.84M	$2.09M	$2.13M	$2.15M	$2.16M	$2.18M	$2.24M	$2.25M
2016-10-02					$1.43M	$1.62M	$1.72M	$1.75M	$1.77M	$1.78M	$1.8M	$1.84M	$1.85M
2016-10-09				$1.8M	$1.57M	$1.62M	$1.73M	$1.76M	$1.78M	$1.78M	$1.8M	$1.85M	$1.86M
2016-10-16			$936.81K	$1.14M	$1.17M	$1.21M	$1.29M	$1.31M	$1.33M	$1.33M	$1.34M	$1.38M	$1.38M
2016-10-23		$520.42K	$679.21K	$733.08K	$756.99K	$780.09K	$830.67K	$845.9K	$856.04K	$857.99K	$865.94K	$889.34K	$893.31K
2016-10-30	$256.57K	$322.32K	$354.5K	$382.62K	$395.1K	$407.16K	$433.56K	$441.5K	$446.8K	$447.81K	$451.96K	$464.18K	$466.25K
All weekly cohorts average CLV													$2.09M

Chart 3.3. Average order value forecast by weekly cohorts.

Next, predictive modelling will be utilized to forecast the missing data, which in this case refers to the revenue expected from later-acquired customer cohorts. For instance, for users whose first purchase occurred in 2016-10-23, we currently have data only for their first week purchase revenue, which amounts to $491 K per customer. However, the revenue for the subsequent weeks remains unclear. The previously calculated Cumulative Growth Percentage (Growth%) will be utilized, and predictions for all 11 future weeks' values will be made. For example, for this cohort, the expected revenue for week 1 can be calculated as $256,57 K x (1 + 27,5%) = $322,32 K, and for week 2 as $322,32 K x (1 + 11.06%) = $354,5 K. By using the average cumulative growth for each week, it can be projected that based on the initial value of $256,57 K, the revenue for week 12 is expected to be $466,25 M.

Utilizing predictive modelling to compute a CLV of $2,09 M showcases a data-driven and precise method for determining customer lifetime value. This approach leverages historical data, patterns, and trends identified through cohort analysis to offer a more accurate estimation of a customer's expected spending over their lifetime with the company.

4. Predictive modelling of probability of purchase

In the dynamic landscape of e-commerce and digital marketing, understanding and predicting consumer behaviour are paramount to the success of businesses. The ability to forecast the likelihood of a customer making a purchase based on their online interactions and acquisition sources holds immense value in optimizing marketing strategies, enhancing customer engagement, and ultimately driving revenue growth.

The logistic regression modelling techniques will be used to predict purchase probabilities using a comprehensive company dataset. The dataset spans from August to October 2016, encompassing 243,297 records and 55 attributes of marketing analytics data, providing a rich tapestry of customer behaviour insights.

The significance of predicting purchase probabilities lies in its direct application to real-world business scenarios:

- **Optimizing Marketing Strategies**: By accurately predicting the likelihood of a customer making a purchase, businesses can tailor their marketing efforts more effectively. This includes personalized messaging, targeted promotions, and optimized ad placements to maximize conversion rates.
- **Resource Allocation**: Understanding purchase probabilities enables businesses to allocate resources efficiently. By focusing on high-probability customers, organizations can prioritize budget allocations towards segments more likely to convert, thereby optimizing ROI on marketing expenditures.
- **Customer Segmentation and Personalization**: Predictive models facilitate the segmentation of customers based on their propensity to purchase. This segmentation enables tailored experiences, such as personalized product recommendations and customized communication strategies.

- **Insights for Strategic Decision-Making**: Predictive analytics on purchase probabilities offer valuable insights for strategic decision-making. By key drivers of purchase intent identification, businesses can refine product offerings, improve website usability, and optimize the customer journey.

The Google Analytics dataset, with its wealth of customer behaviour and acquisition source data, presents a unique opportunity to delve into these predictive modelling techniques. Through exploratory data analysis (EDA), data preprocessing, handling missing values and outliers' management, logistic regression modelling, this project aims to develop a robust framework for predicting purchase probabilities.

Python will serve as the primary tool for implementing this project, leveraging libraries such as pandas, NumPy, scikit-learn, and seaborn for data manipulation, modelling, and visualization.

Data Preparation:

The initial dataset underwent rigorous preprocessing to ensure data quality and model robustness. Missing values were addressed by first identifying columns with over 50% missing values, which were subsequently dropped from the analysis due to significant data loss. For remaining missing values in categorical variables, the mode was used for imputation, while numerical variables were imputed using the median or mean, depending on the distribution and nature of the data.

Outlier Management:

Outliers, defined as data points lying far from most observations, were a critical consideration in this project. After evaluating the impact of outliers on the model's performance, a strategic decision was made to retain outliers even if they constituted up to 19% of the dataset. This approach was motivated by the hypothesis that outliers could contain valuable insights, particularly in estimating the probability of purchase, a crucial aspect of the predictive model's objective.

Feature Selection and Multicollinearity:

To optimize the feature set for modelling, each feature's correlation with the target variable was assessed. Additionally, multicollinearity, the presence of highly correlated independent variables, was investigated. In instances where multicollinearity was detected, only one variable from the correlated set was retained—specifically, the variable demonstrating the highest correlation with the target variable. This streamlined approach not only simplified the model but also mitigated issues of redundancy and potential overfitting.

Logistic Regression Modelling and Evaluation:

1. Data Splitting: The dataset was divided into training and testing subsets to assess the performance of the logistic regression model. The training dataset was used to fit the model, while the testing dataset was kept aside to evaluate its predictive capabilities on unseen data.

2. Feature Selection Based on Statistical Significance: Initially, all available features were included in the logistic regression model. However, to refine the model and improve its interpretability, a feature selection process was undertaken. Features with p-values greater than 0.05 were deemed less statistically significant and hence excluded from the final model. This step aimed to enhance the model's predictive power by focusing only on the most impactful features.

3. Adjusting Model Cut-off Threshold Using ROC Curve Analysis: The Receiver Operating Characteristic (ROC) curve was utilized to analyse the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) across various threshold values. By assessing the ROC curve, a specific cut-off threshold of 0.15 was selected to optimize the model's performance, balancing precision and recall based on the project's objectives.

4. Model Performance Metrics: After tuning the model with the selected features and cut-off threshold, key performance metrics were computed. The logistic regression model achieved the following results:

- Precision: 37.58%
- Accuracy: 98.57%
- Recall (Sensitivity/True Positive Rate): 33.6%

Interpreting Model Parameters

Precision: This metric indicates the proportion of predicted positive cases (purchase probability) that were true positives. In this context, the model correctly identified 37.58% of the predicted purchases among all positive predictions.

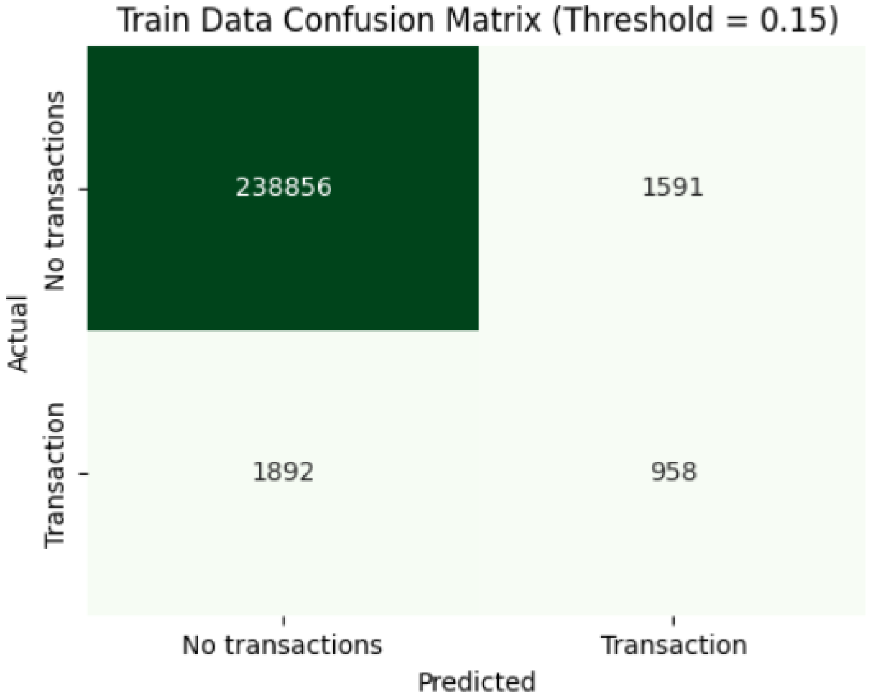


Chart 4.1. Logistic regression model confusion matrix

Accuracy: This metric measures the overall correctness of the model across all predictions. The high accuracy of 98.57% suggests that the model performed well in classifying both positive and negative cases.

Recall (Sensitivity/True Positive Rate): This metric quantifies the model's ability to correctly identify actual positive cases (purchases) out of all true positive cases in the dataset. A recall of 33.6% indicates that the model captured about one-third of all actual purchases.

True Positive Rate (Sensitivity/Recall) and False Positive Rate

True Positive Rate (Sensitivity/Recall): This value (0.3361 or 33.6%) signifies the proportion of actual positive cases (purchases) that were correctly identified by the model.

False Positive Rate: This rate (0.0066 or 0.66%) indicates the proportion of actual negative cases (non-purchases) that were incorrectly classified as positive by the model.

To interpret the probabilities of purchase associated with significant features identified in your logistic regression model, we can analyse the impact of each feature on the estimated probability of purchase. Here's a breakdown of the probabilities associated with the significant features:

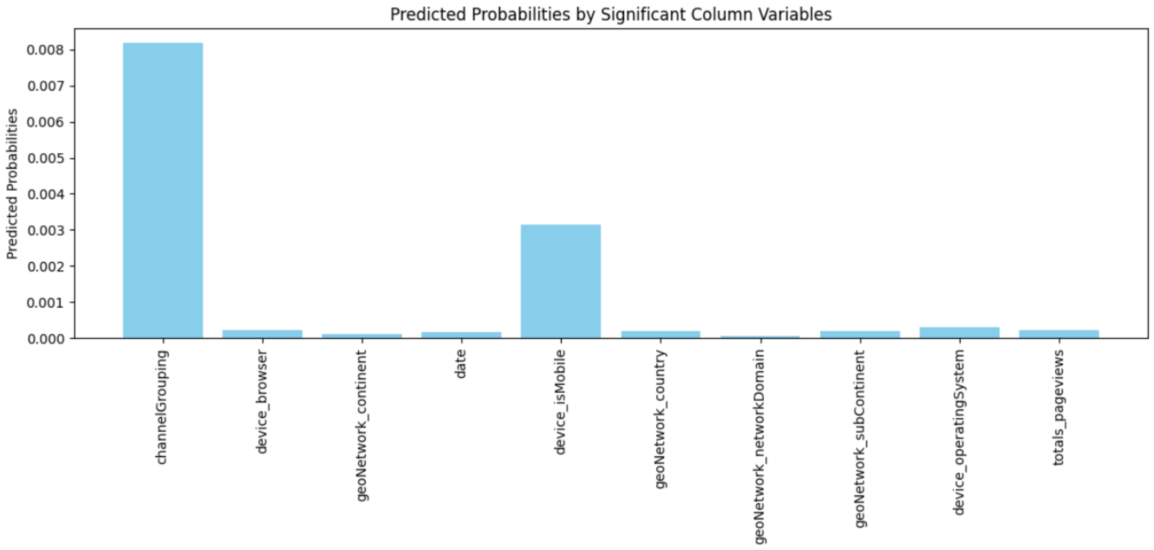


Chart 4.2. Probability of purchase by significant variables

channelGrouping (0.0082):

This probability (0.0082 or 0.82%) indicates that users from different channel groupings (such as Organic Search, Paid Search, Direct, etc.) have varying likelihoods of making a purchase. A higher probability suggests that certain channel groupings may be more effective in driving purchases.

device_browser (0.0002): The probability associated with the user's web browser (0.0002 or 0.02%) suggests that certain browsers used by visitors may slightly influence purchase behaviour. However, the impact appears to be relatively minor compared to other factors.

geoNetwork_continent (0.0001): The continent from which the user originates contributes to a very low probability (0.0001 or 0.01%) of purchase. This implies that geographical location at the continent level may have limited influence on purchase decisions in this context.

date (0.0002): The specific date of the visit appears to influence the purchase probability to a small extent (0.0002 or 0.02%). This suggests that temporal factors, such as seasonality or special events, might impact user behavior and purchase likelihood.

geoNetwork_country (0.0032): Users from different countries exhibit varying probabilities of making a purchase, with a probability of 0.0032 (0.32%). This highlights the role of country-specific factors, such as economic conditions or cultural preferences, in driving purchase behaviour.

geoNetwork_networkDomain (0.0002): The domain of the user's network connection contributes minimally (0.0002 or 0.02%) to the probability of purchase. Network domain may reflect organizational affiliations or service providers, but its impact on purchasing decisions appears limited in this analysis.

geoNetwork_subContinent (0.0001): Sub-continental regions also play a marginal role (0.0001 or 0.01%) in influencing purchase probabilities. This suggests that broader geographical categorizations have relatively minor impacts compared to more specific country-level data.

device_operatingSystem (0.0002): The user's operating system contributes slightly to the purchase probability (0.0002 or 0.02%), indicating that device-related factors may influence user engagement and conversion rates.

totals_pageviews (0.0003): The number of pageviews generated during the visit correlates with a small increase in the probability of purchase (0.0003 or 0.03%). This implies that user engagement, as measured by pageviews, may be a predictor of purchase intent.

These probabilities highlight the nuanced influences of various features on the likelihood of purchase within the analysed dataset. Factors such as the user's origin (country, continent), browsing behaviour (channel grouping, browser), and temporal context (date) all contribute to shaping purchase probabilities. Understanding these insights can inform targeted marketing strategies, user experience optimizations, and personalized content delivery to maximize conversion rates and revenue generation based on specific user segments and behaviours.

Further probability of purchase evaluation strategies:

1. Address Class Imbalance

Resampling Techniques: Consider using resampling methods such as oversampling the minority class (transactions) or under sampling the majority class (non-transactions) to balance the dataset. Techniques like Synthetic Minority Over-sampling Technique (SMOTE) can be effective in generating synthetic samples of the minority class to achieve a more balanced distribution.

2. Feature Engineering with Clustering

- WOE (Weight of Evidence) Analysis: Explore feature engineering techniques like WOE analysis to transform categorical variables into informative features that capture the relationship between each feature and the target variable (purchase/non-purchase). This can enhance the discriminatory power of your model.

- Cluster Analysis: Utilize clustering algorithms (e.g., K-means clustering) to identify meaningful segments within your dataset based on user behaviour or attributes. By incorporating cluster labels as additional features, you can capture hidden patterns and improve the model's predictive performance.

3. Explore Alternative Models

- Decision Tree (and Ensemble Methods): Implement decision tree-based algorithms such as Random Forests or Gradient Boosting Machines (GBM). Decision trees are robust to class imbalance and can capture nonlinear relationships and interactions between features effectively.
- K-Nearest Neighbours (KNN): Experiment with KNN, a non-parametric algorithm that makes predictions based on the similarity of data points. KNN can be particularly useful for identifying localized patterns and handling imbalanced datasets.

4. Model Evaluation and Parameter Tuning:

- Cross-Validation: Employ cross-validation techniques (e.g., k-fold cross-validation) to assess the generalizability of your model and mitigate overfitting.
- Hyperparameter Tuning: Fine-tune model hyperparameters (e.g., tree depth, number of neighbours in KNN) using grid search or randomized search to optimize model performance.

5. Ensemble Learning:

Ensemble Methods: Combine multiple models (e.g., Random Forest, Gradient Boosting) using ensemble techniques like stacking or boosting. Ensemble learning can further enhance predictive accuracy by leveraging the strengths of diverse models.

Conclusions

In conclusion, this comprehensive analysis underscores the importance of understanding customer behaviour and preferences to drive growth and enhance customer satisfaction. Most of the revenue is attributed to **referral, direct** and **organic search channels**, with a significant contribution from desktop users, particularly within the **US market**.

Utilizing predictive modelling to compute a Customer Lifetime Value (CLV) of $2.09 million showcases a data-driven and precise method for determining customer lifetime value. This approach leverages historical data, patterns, and trends identified through cohort analysis to offer a more accurate estimation of a customer's expected spending over their lifetime with the company.

With 54.4% ($214.48 billion) of revenue attributed to **Best Customers** and **Customers Needing Attention**, it's crucial to not overlook other customer segments, especially the 34.9% ($156.16 billion) categorized as **At Risk** and **Loyal Customers**. The company should focus on retention and re-engagement strategies to address these segments.

Analysing Customer Lifetime Value, sales trends, and marketing spendings provides valuable insights into the business's overall performance. Additionally, probability of purchase analysis might help in anticipating customer behaviour and optimizing resource allocation. By assessing the likelihood of a customer making a purchase based on historical patterns and predictive factors, businesses can prioritize their marketing efforts more efficiently.

Understanding the probability of purchase can guide personalized marketing strategies, such as sending timely and relevant promotions or incentives to customers who are most likely to buy. Moreover, this analysis can inform inventory management decisions, ensuring that the right products are available when demand is anticipated to be high.

The logistic regression model, refined through feature selection and threshold optimization, demonstrated strong overall accuracy but with modest precision and recall specific to identifying purchases. These insights provide valuable guidance for further model refinement and strategic decision-making in leveraging predictive analytics for estimating purchase probabilities in a retail setting.

Finally, by applying these insights and recommendations, businesses can optimize marketing strategies, improve customer targeting and engagement, and ultimately drive sustainable revenue growth and long-term customer loyalty.

Recommendations

Based on the findings, the following recommendations are proposed:

- **Focus on High-Value Segments**: Prioritize marketing and retention efforts on Loyal Customers and Customers Needing Attention, as they collectively contribute significantly to the company's revenue (54.4%, $214.48 billion). Tailor marketing messages and offers specifically to these segments to maximize their engagement and lifetime value. Implement personalized campaigns and incentives based on their preferences and purchasing behaviour to foster loyalty and increase repeat purchases.
- **Engage with At Risk, Best Customers, and Hibernating Segments**: Address the At Risk, Best Customers, and Hibernating segments which collectively contribute approximately 34.9% ($156.16 billion) of revenue. Initiate proactive engagement strategies such as surveys, feedback forms, or personalized communications to understand their concerns and needs. Develop targeted solutions to re-engage these segments, emphasizing the benefits of continued loyalty to the brand. Leverage customer insights to tailor promotions and incentives that resonate with their preferences and motivations.
- **Enhance Probability of Purchase Estimation**: Improve the accuracy of probability of purchase estimation using advanced strategies and techniques:
    - Address class imbalance by employing techniques like SMOTE (Synthetic Minority Over-sampling Technique) for oversampling or exploring under-sampling methods to balance the dataset.
    - Conduct feature engineering using methods like Weight of Evidence (WOE) analysis and cluster-based segmentation to enrich the dataset with predictive features.
    - Experiment with alternative machine learning models such as decision trees (e.g., Random Forest) and K-Nearest Neighbors (KNN) to capture complex patterns in customer behavior.
    - Ensure model interpretability by leveraging SHAP (SHapley Additive exPlanations) values and collaborating with domain experts to gain business insights and validate model outputs.