Customers segmentation and probability to purchase modelling

Denis Senin
2024-05-07

Agenda

- Introduction & objective
- Data source & data preparation
- Exploratory analysis
- Main analysis
    - RFM customer segmentation
    - Customer lifetime value (CLV) Analysis
- Probability to purchase prediction
    - Logistic regression
- Key Insights & recommendations

Introduction

Objective : Enhance customer experience, elevate satisfaction levels, boost revenue

Methodology: RFM Customers Segmentation, Customer Lifetime Value (CLV) using BigQuery and Looker Studio, Logistic Regression using Python.

Target Audience: Transacted and non-transacted electronics company portal customers

Goal: Provide actionable insights to ensure growth and improve operations

Introduction to Data Source

E-Commerce Data Source for 2016 Aug 1 – 2017 July 31.

Dataset information: visitors data including geographical, technological, marketing channels usage, transactional revenue

Dataset size: ~243k rows, 55 attributes

Selected Time Coverage: marketing analytics dataset (2016 Aug 1 - 2016 Oct 31)

Data Preparation

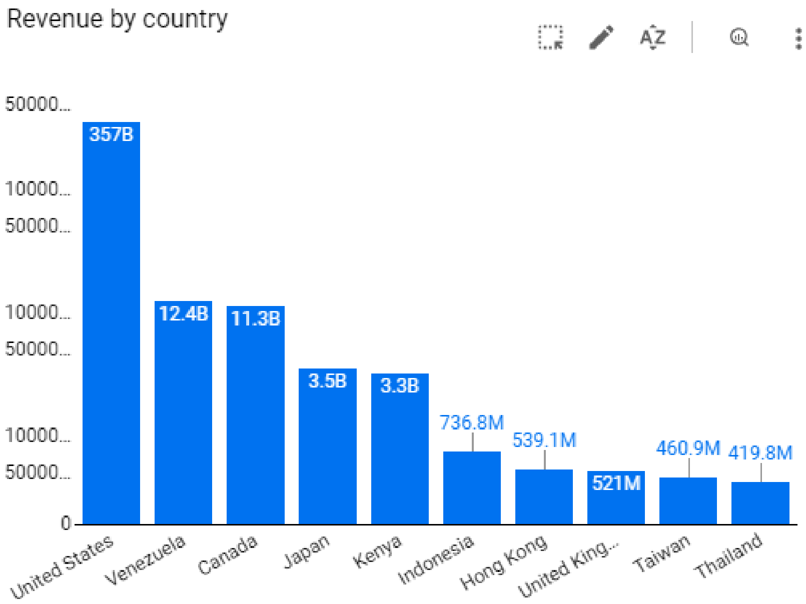
Main analysis	Probability of purchase analysis	Features selected
Feature Engineering: Checking the revenue values of customer transactions data, creating a necessary transactional field for further analysis.	**Handling Missing Values and Duplicates:** checking for missing and duplicated data to fix or replace it by median or mode.	24
	Outlier Identification and Management: checking and identifying outliers to manage them and normalize the variables distribution.	
	Variables encoding: categorical and boolean variables encoding to make them suitable for logistic regression modelling.	
	Feature Selection Based on Correlation and Multicollinearity	20
	Feature Selection by Significance: variables have been chosen for modeling based on their significance.	14 (10)

Exploratory Analysis (1)

Key Metrics:

- ~200k unique visitors, ~2k unique transactions.
- $394.03 billion in revenue, $138,2 millions average transactions value

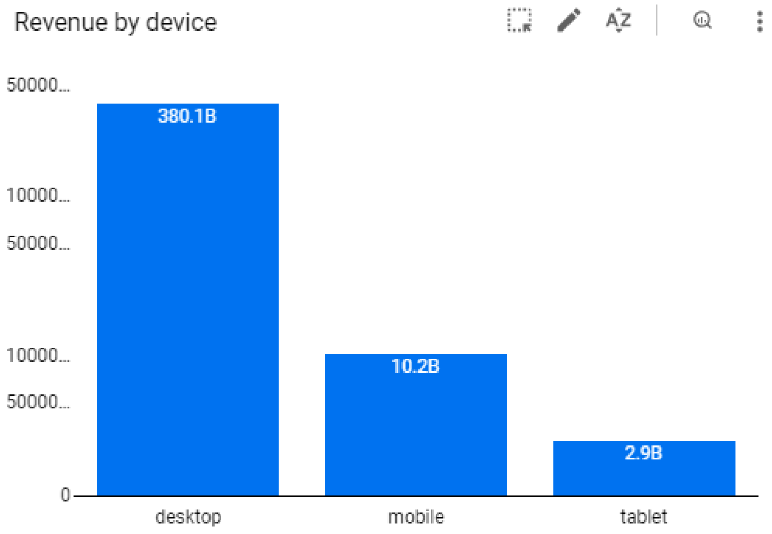
Exploratory Analysis (2)



Key Metrics:

Customers from USA generates ~ 91% of the revenue, compared to other countries only ~9%.

Exploratory Analysis (3)

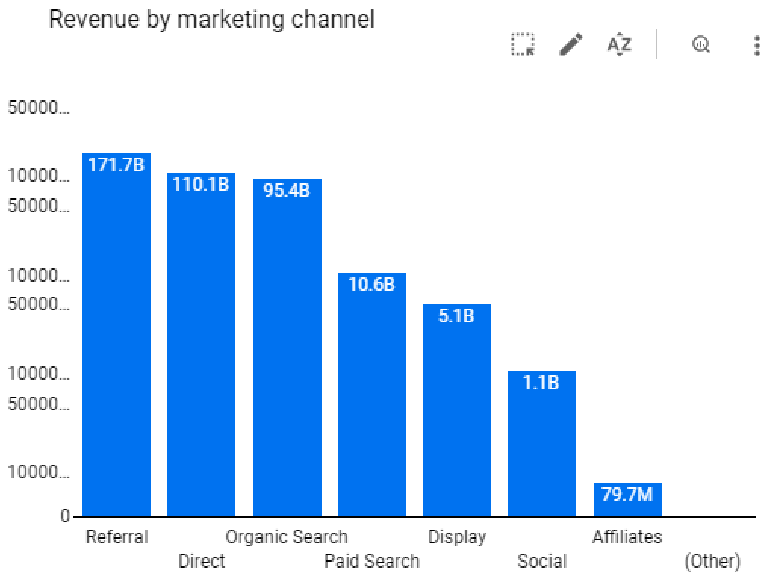


Key Metrics:

Desktop user generates ~ 97% of the revenue,

compared to mobile and tablet only ~3%.

Exploratory Analysis (4)



Key Metrics:

Most of revenue is generated from Referral,

Direct and Organic Search marketing channels

Exploratory Analysis (5)

RFM segment	# of custo...	% of custo...	Revenue
Loyal Customers	529	21%	$130.18B
Customers Needing Attention	708	28%	$84.3B
At Risk	168	7%	$71.87B
Best Customers	127	5%	$65.93B
Hibernating	348	14%	$22.27B
Cant Lose Them	25	1%	$8.66B
Recent Customers	186	7%	$3.04B
Promising	171	7%	$2.83B
About to Sleep	146	6%	$2.57B
Lost Customers	133	5%	$2.38B

Key Metrics:

Best customers and **Customers Needing Attention** generates most of the revenue across customers segments

Exploratory Analysis (2)

Key Metrics:

- Transacted visitors show high engagement rate in average - 34 pagevies vs non-transacted – only 4.

RFM Analysis

- Customer Segmentation based on Recency, Frequency, Monetary values:
  - Non-transacted – **engagement**: visit recency, number of visits, number of pageviews
  - Transacted – **purchase**: transaction recency, number of transactions, average revenue
- Customer distribution across segments
- Revenue distribution across segments

RFM Analysis – Transacted customers

Key Insights

- Top Segments: **Best Customers** and **At Risk**
- Segments **Cant Loose Them**, **Customers Needing Attention** need extra care
- Segment **Loyal customers** need to be nurtured and up-selling and cross-selling strategies should be applied
- Need to adjust the marketing strategies for segments.

Customers Revenue Distribution across segments

	RFM segment	# of visitors	% of visitors	Recency (Average…	Frequency (Av…	Monetary (Averag…
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…	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
1…	Recent Customers	186	7%	13.4	1	$16.34M

RFM Analysis – Non-transacted customers

Key Insights

- Top Engaged Segments: **At Risk, Cant Loose Them**, **Best Customers**, **Loyal Customers**.
- Segments **Customers Needing Attention** and **Hibernating** need extra care.
- Need to improve user experience and maintain re-engagement strategies for segments with higher monetary (engagement) metrics).

Customers Engagement rate by segments

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

Customer Lifetime Value (CLV) Analysis

- Monthly Customer Lifetime value for 2016 Aug 1 – 2016 Oct 31
- Cohort analysis and Predictive Modeling results shows Overall CLV = $2,09 M

CLV Analysis

Key Insights

- All Customers CLV : $2,09 M
- Last 3 weekly cohorts shows higher predictive values

Predictive Modeling of CLV calculation

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

Probability of purchase analysis

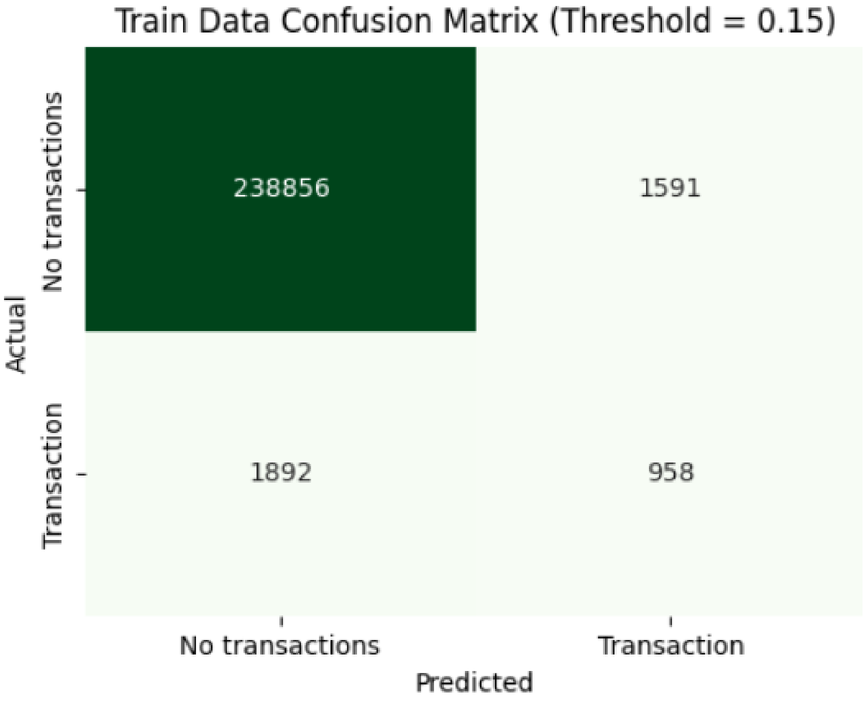
- Selecting significant attributes for predictive modelling
- Logistic regression model creation
- Probability of purchase estimation

Probability of purchase model creation

Key Insights

- Model created with precision – 37,58%, accuracy - 98.57%, recall -33.6%.
- True Positive Rate (Sensitivity/Recall): 0.3361, False Positive Rate: 0.0066
- Need to balance dataset to have a better prediction for positive cases
- It is recommended to use alternative methods as KNN and decision tree to compare the model performance

Logistic regression model

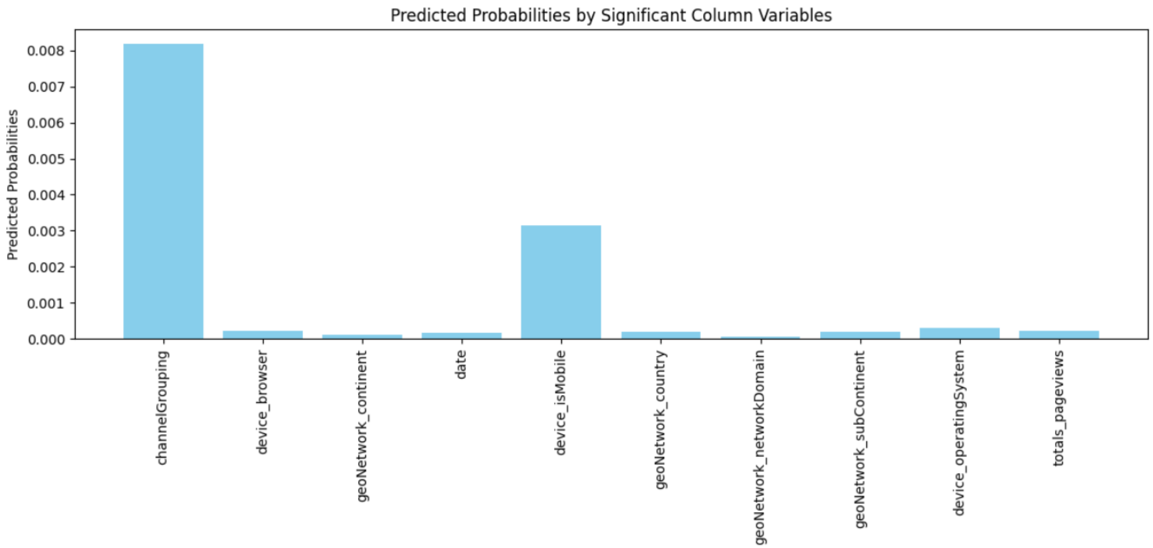


Probability of purchase

Key Insights

- Higher probabilities has variables as marketing channels and devices related data (type)
- Possibilities to scale most significant categories involving feature engineering by they options
- Additional data required for model improvement

Most significant factors predicting probability of purchase



Key insights & Recommendations

Key Insights

1

Referral, direct and organic search channels brings most traffic and revenue

2

USA market generates the most revenue

3

Majority transactions and revenue comes from desktop users

4

Best Customers & **Customers Needing Attention** customers bring 54,4% of all revenue ($214,48 billion).

5

Predictive Customer lifetime value ($2.09 M)

6

The highest impact for purchase probability has marketing channel and device type related data

Recommendations

1

Improve user experience for desktop and mobile users to maximize revenue and number of transactions

2

Focus on retention and re-engagement strategies

3

Focus on High-Value Segments, tailor marketing messages to maximize engagement and generated revenue

4

Engage with **At Risk** and **Loyal Customers** which account is 34,9% of customers ($156,16 billion)

5

Improve probability of purchase estimation by using strategies like balancing dataset, feature engineering with WoE estimation

6

Use alternative models such as decision trees (e.g., Random Forest) and KNN to capture complex patterns.

Thank you!