Group - 4: Soumya Agrawal, Olivia (Ryunghee) Lee, Aniket Patil, Soumith Palreddy

Customer Segmentation and Targeting

Description of project goals

Our project aims to identify the high value customers of an online retail platform and delineate customer persona to inform advertising budget decision-making. We are using the transactions data of a UK-based platform spanning from December 2009 to December 2011 to achieve our goal.

Importance of the Problem

The global online retail market is valued at 4.75 trillion USD and is expected to grow at a CAGR of 9.4 percent for the next 7 years. Online retail businesses spend about 6 to 20% of their revenue on marketing. Driving more revenue per dollar spent on marketing is ever more important for bootstrapped startups to retain and engage their customers. Undoubtedly, making it one of the more important problems to solve.

The "Online Retail UCI" dataset has ~1 million records of transactional sales data which spans across 38 countries and includes both wholesale and retail customers making it a robust dataset to work and generate insights.

Exploratory Analysis

Here are some key stats and data cleaning process on the dataset:

1. Size of the Dataset: 1067371 transactions (rows) and 8 columns
2. 23% of the rows don't have a "CustomerID" and the rows are dropped
3. Customer Count : 5942
4. # of Purchases: 53628
5. Time Period: 1st Dec, 2009 to 9th Dec, 2011 (~24 months)
6. 91% of the data is from UK
7. Data Outliers have been handled by capping the values at 1.5 times the inter-quartile range from the 1st and 3rd quartile
8. Rows having negative values for "quantity" (22950 rows) and "unit price"(5 rows) have been dropped.

Some key insights from the dataset:

1. Seasonality:
    a. Nov, Oct and Dec record the highest sales while its lowest in Feb – Q'4 records the maximum sales *(Fig1 & 2)*
    b. Sales are highest on Thursday and the lease on Saturday
    c. Afternoons record the highest sales and evenings are the lowest
2. Repeat Customers:
    a. Most of the repeat customers purchased 13 or fewer times in the dataset
    b. The majority of repeat customers tend to make a purchase every 12 to 60 days

Cohort Analysis: Top 25 percent of the customers (fig 3)

The main characteristics:

1. There are some empty spaces on the 2010 December Cohort. This means the top 25 % of the customers in the 2010 December cohort did not purchase during June 2011, September 2011, and October 2011.
2. Since it's showing the top 25% of the customers, the overall percentage went up higher than all of the customers.
3. Compared to the heatmap of all customers, the top 25% of customers are actively purchasing in the cohorts of 2011 February till 2011 October. (more orange colors: higher percentage)

Solution and Insights

We approached the problem in two ways - (a) A point-in-time snapshot of customer persona through RFM analysis, and (b) A running 3-month CLTV model to continuously track potential high-value transacting customers.

RFM Analysis:

RFM stands for Recency, Frequency and Monetary; these customer level metrics are often used to segment customers in the retail industry and based on the RFM score the next best targeting action for the customer is decided.

Segment	# of Customers	Description	Next Best Action
Loyal	1100	Customers who buy most often from the store	Loyalty Programs, Free Shipping
Whales	370	Customers who generate high revenue (Highest Monetary Value)	Premium offers, Subscription tiers, Luxury Products
Promising - Faithful	199	Customers who visit Often but don't spend a lot	Incentives tied to spending thresholds and product recommendations

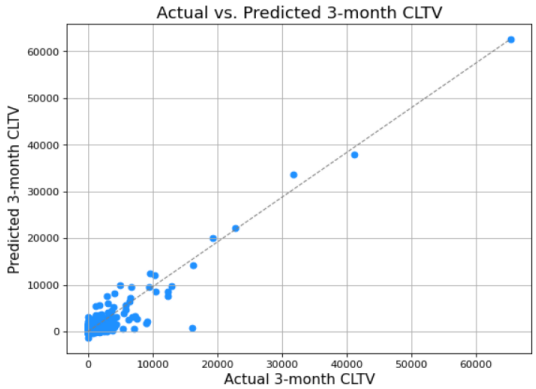
3-month CLTV model -

A 3-month CLTV (Customer Lifetime Value) measures the total income a business can expect to bring in from a client over the course of the next 3 months. We have extracted the below features that span over 24 months in 3 month time-steps:

1. Sum of sales
2. Count of Sales
3. Average of Sales

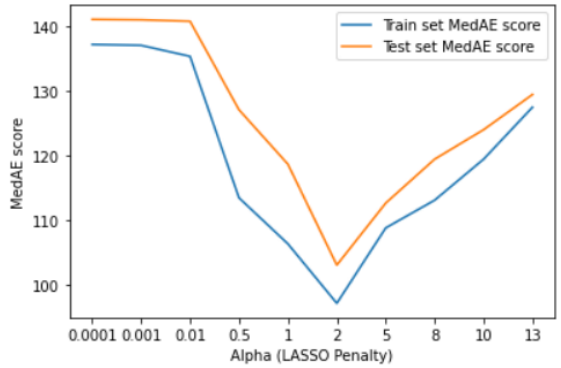
This totals to 24 features (8 quarters for each feature). The intuition being that our model can identify trends in the time-series for each customer. This poses the obvious problem of multicollinearity among the features which is handled by our choice of model.

As a baseline model, we fit an OLS multiple linear regression model. We expected a LASSO model to outperform the baseline, since it can zero out the irrelevant features given a large enough alpha (penalty) and tackle multicollinearity. The results from the baseline model are summarized below-



Baseline Train set MedAE: $ 137.13
Baseline Test set MedAE: $ 141.01
Range of target: $0 to $65,000

The LASSO model improved upon the baseline by 8.6% on the test set -



Train set MedAE for alpha=2 is $ 116.65
Test set MedAE for alpha=2 is $ 128.87
Range of target: $ 0 to $ 65,000

The figure shows that Median Absolute Error (MedAE) is the least at alpha=2. After that, important features get penalized and we lose accuracy.

Since we get comparable results for train and test sets and cross-validation results look good, we are confident that our model is not overfitting.

	coef	std err	t	P>|t|	[0.025	0.975]
sales_avg_M_2	0.0231	0.037	0.616	0.538	-0.050	0.097
sales_avg_M_3	0.4059	0.053	7.710	0.000	0.303	0.509
sales_avg_M_4	-0.4475	0.023	-19.257	0.000	-0.493	-0.402
sales_avg_M_5	-0.3738	0.038	-9.754	0.000	-0.449	-0.299
sales_avg_M_6	-0.1965	0.045	-4.320	0.000	-0.286	-0.107
sales_avg_M_7	-0.0341	0.040	-0.849	0.396	-0.113	0.045
sales_avg_M_8	-0.0704	0.041	-1.712	0.087	-0.151	0.010
sales_avg_M_9	0.5175	0.075	6.864	0.000	0.370	0.665
sales_count_M_2	104.7772	11.244	9.318	0.000	82.731	126.823
sales_count_M_3	33.1731	11.281	2.941	0.003	11.055	55.291
sales_count_M_4	0	13.186	0	1.000	-25.854	25.854
sales_count_M_5	-11.3120	8.186	-1.382	0.167	-27.361	4.737
sales_count_M_6	-65.0468	11.130	-5.844	0.000	-86.868	-43.226
sales_count_M_7	22.3619	12.501	1.789	0.074	-2.148	46.872
sales_count_M_8	-15.2852	12.119	-1.261	0.207	-39.046	8.475
sales_count_M_9	-50.4404	21.874	-2.306	0.021	-93.327	-7.553
sales_sum_M_2	0.1289	0.011	11.366	0.000	0.107	0.151
sales_sum_M_3	-0.1024	0.020	-5.128	0.000	-0.142	-0.063
sales_sum_M_4	0.4556	0.021	21.231	0.000	0.413	0.498
sales_sum_M_5	0.4009	0.014	28.666	0.000	0.373	0.428
sales_sum_M_6	0.1457	0.022	6.528	0.000	0.102	0.189
sales_sum_M_7	-0.1382	0.020	-7.009	0.000	-0.177	-0.100
sales_sum_M_8	0.0624	0.018	3.557	0.000	0.028	0.097
sales_sum_M_9	-0.2056	0.049	-4.178	0.000	-0.302	-0.109

The LASSO model summary to the left shows important features. Coefficient for last 4th quarter sales count is zeroed by LASSO signifying it has no impact on 3-month CLTV. Also, many sales count features have t-value < 2, rendering them statistically insignificant as the 97.5% C.I. includes 0.

Surprisingly, the average sales of the last quarter is not statistically significant for the target. This could be because many customers make the first purchase in response to a marketing campaign and then churn out.

We propose to fetch high-value customers from this 3-month CLTV LASSO model and match them with RFM analysis personas, to maximize return on advertising budget.

Figures:

Fig1:

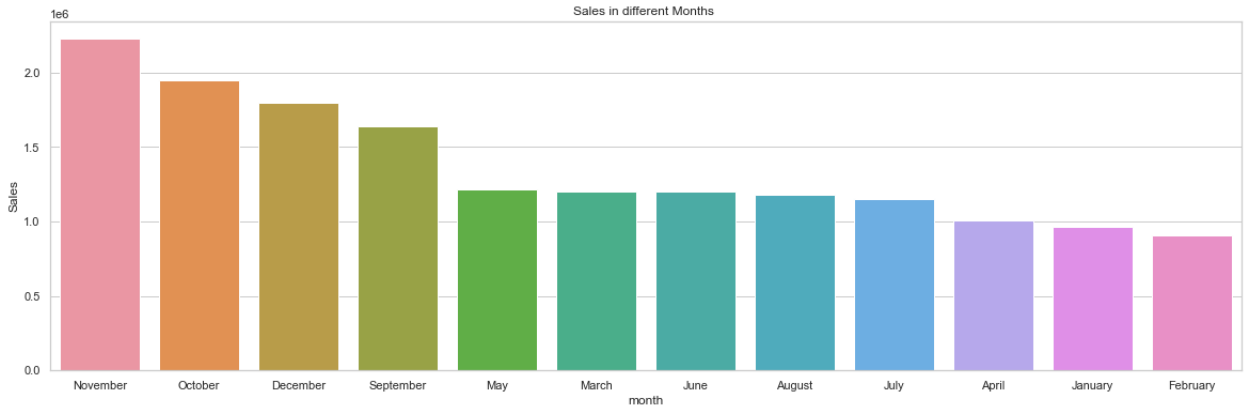


Fig2:

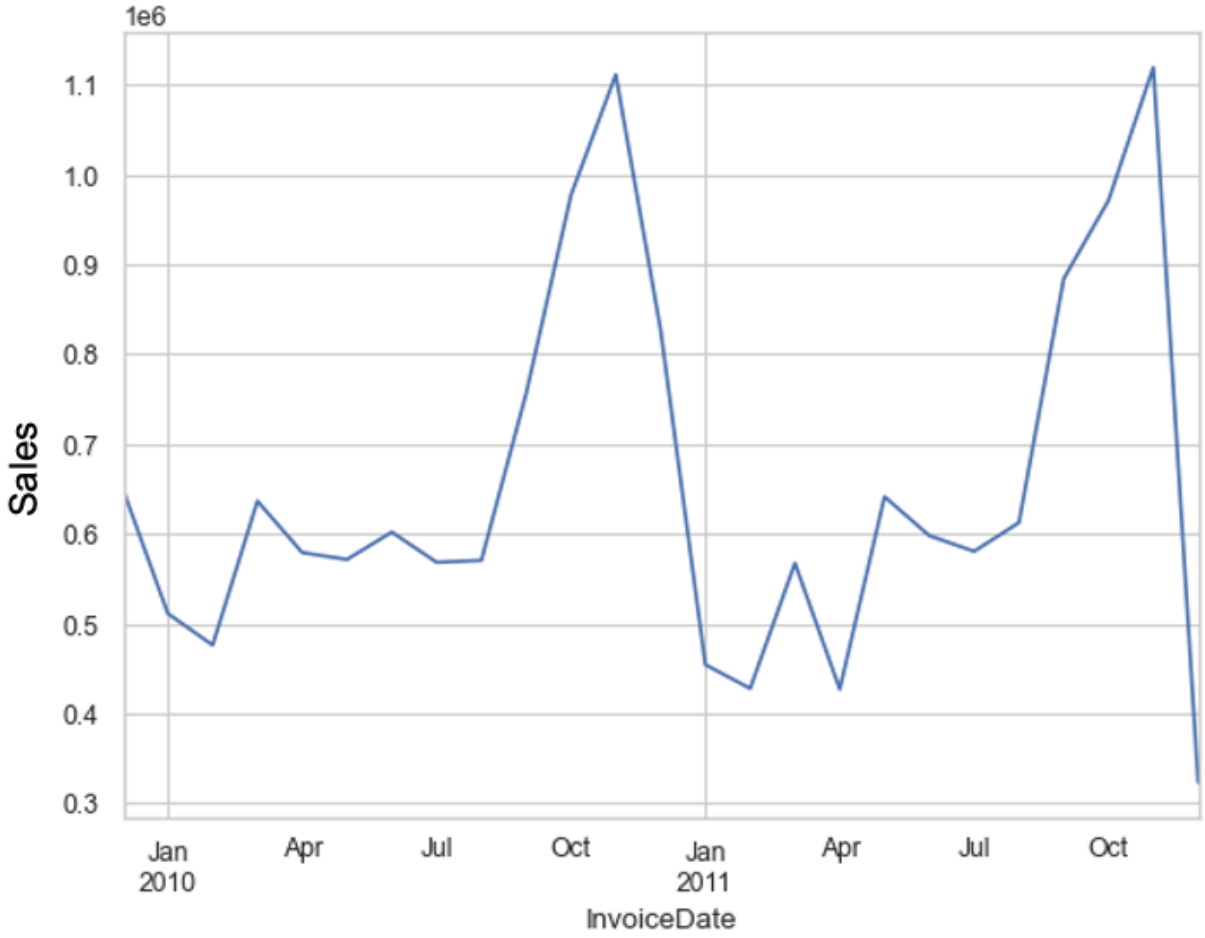
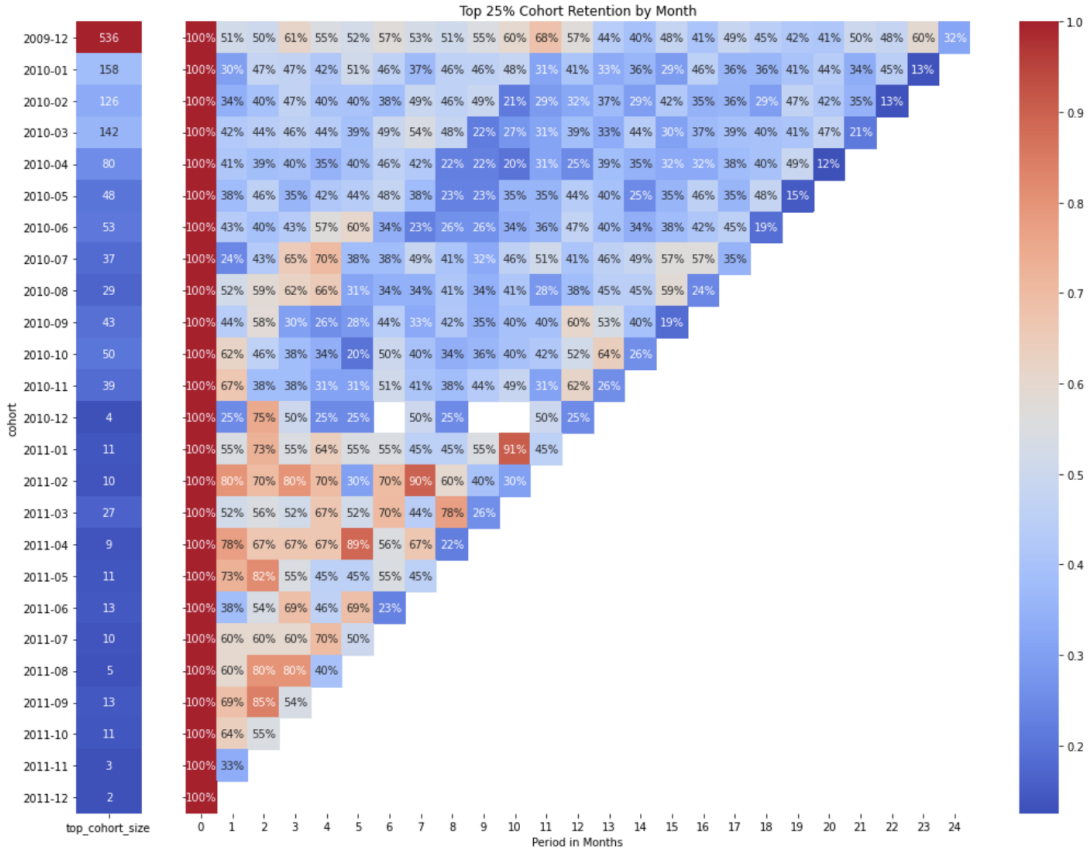


Fig3:



Heat map for TOP 25 percent of the customers