[draft for medium blogpost]

**Find distinct customer segments based on their purchasing behavior using unsupervised learning. Starbucks challenge!**

This dataset is provided by Starbucks and a hard nut to crack. I had to go all the way to find clear patterns: Complex cleaning, intense preprocessing, fine-tuned dimensionality reduction with PCA and then a final clustering exercise with k-means. I'll gladly give you an overview, the details are provided in my github repository.

**Introduction**

The data contains 3 sets of transaction, demographic and offer data. It is simulated and mimics the behavior for 17'000 customers on the Starbucks rewards mobile app over the course of 29 days. Once every few days, Starbucks sends out an offer to users of the mobile app. An offer can come in three types

1. a discount offer (discount)
2. a buy-one-get-one-free offer (BOGO)
3. merely an advertisement for a drink (info)

Every offer has its own characteristics (e.g. amount to be spent to get the reward) among which one of the most important is the duration, denoting the time by which the spending threshold as to be reached to qualify for the reward. (This is for discounts and BOGOs only).

The goal of the project is to determine which demographic groups respond best to which offer type.

I will present my findings in 5 sections:

1) Cleaning / Feature Engineering

2) General Offer / Offer Type Comparision

3) PCA and Clustering

4a) Segment Analysis Based on Customer Behavior

4b) Conclusions on Demographic Groups

5) Learnings

Before we start let me say the following:

As can be seen from the following simple plot the general strategy seems to work. A release of offers is followed by viewings of these offers, a spike in transactions and eventually, if the spending threshold is met, with offer completions that lead to a reward.

[graph 1]

But there's two things making this project a *real challenge*:

1) There is no apparent experimental setup and we have no control over the variables: Not all users receive the same offers / offer types or the same amount of offers. During certain weeks some customers may receive multiple offers at the same time, while others don't receive any at all. Also, the different offers have different characteristics making their successful completion more or less likely.

2) The transaction data needs extensive and pretty complex preparation. The transactions a customer makes are not linked to any offers he has received. And even when an offer is sent to a customer this doen't mean he has viewed it and hence is influenced by it. Also, we have to make sure that we only count the valid offer completions, that's those that occur within the period of validity.

**1) Cleaning / Feature Engineering**

I will skip the basic cleaning and dive right into the problem mentioned above. I solved this in two steps:

1. Flagging of all events that could be linked to an offer starting from the moment it was received until the period of validity ends (with help of the offer's duration feature).
2. Make sure that of those flagged events only those are counted that occur after an acutal viewing by the customer and no longer as to the moment the customer has completed the offer.

A word of warning: I did this with a couple of encapsulated for-loops, a procedure iterating on my computer for about 20 (!) hours on 317000 rows over 10 offer columns... No details here, but the pictures of the transaction data for one random customer illustrate the results:

[graph 2 / 3, p88, including explanation]

This gave me a solid data base for the subsequent steps.

**2) General Offer Type Comparision**

But first let's look at the different offers and their characteristics. Too many uncontrolled variables for a direct comparison. Although there are some general trends for sure (see my EDA notebook if you want) I could not find meaningful results that would really help me to solve the problem.

There is a somewhat more robust trend to see, if we aggregate the offers by type and compare some key metrics:

[table 4, agg offer types]

First let me explain the view-to-complete rate (vtc rate), the most important metric for me in this project. I have calculated it as the amount of offers that a customer has actually viewed and completed within their respective period of validity divided by the amount of all offers the customer has viewed. It can range from one (all viewed offers have also been completed by the respective customer, meaning he reacts well) to zero (none of the viewed offers are viewed, meaning the customer does not react well).

So back to the comparison of the offer types: It is clear to see, that the average vtc rate is significantly higher for discounts than for BOGOs. This is even though the monetary rewards for BOGOs are much higher. (Note: The higher average difficulty (= spending threshold) of discounts is offset by their longer duration: I calculated this as the 'relative difficulty', the mean amount to be spent per day to reach the completion threshold before the offer ends.)

So, this is interesting - why would a company want to send out expensive BOGO offers when it can have a better activation with the cheaper discounts?

Now we should ask ourselves:

- Have the two offer types been viewed by approximately the same groups of customers?

- Are there certain customer segments that are responsible for the differences or is this a general trend?

**3) PCA and Clustering**

The dataframe I worked with now had a row for every customer and a lot of constructed features describing his overall purchasing behaviour and the behaviour concerning each offer type specifically. I also added the non-promo condition as a fourth offer type.

As our main goal is to find out if the purchasing pattern differ for different demographic groups, I removed all demographic features (including the duration of membership) from the set. I think this is important and something that often gets' done wrong.

There is a lot of experimentation documented in my repository considering pre-processing, and fine-tuning of PCA and clustering. To make a long (and intersting) story short: I log-transformed my data, removed some outliers, scaled it to a range of 0,1, applied a PCA to reduce the feature space to 2 dimensions and the clustered with k-means to have 12 segments.

A note to outlier removal: yep, I removed approx 5% of the customers. But I really prefer to have a more general model than to make compromises for customers that behave so special that your proposed treatments don't apply to them anyway.

See 2 plots showing the results of this procedure:

[graph 4 barplot, graph 5 segments in bi-plot, with explanations]

**4a) Segment Analysis Based on Customer Behavior**

Remember that the 12 segments have been based on purchasing behaviour only. The cool thing here is that they let me control for the exposure of the respective customers to different offer types. The findings are listed below in somewhat simplified form. (For simplicity's sake I wont mention info offers in the segment descriptions, but will add a remark in the end)

First group of segments - customers that have viewed BOGO offers and discounts in a similar share:

- Seg 1 (15.4% of total customers, biggest segment): Customers with highest spending / net revenue, react very well to discounts (mean view-to-complete rate of 0.9) and nearly as good to BOGOs (0.8)

- Seg 2 (14%, 2nd biggest segment): Customes with 2nd higest spending / net revenue, same vtc rate on discounts as Seg 1 but BOGO vtc rate drops to 0.6

- Seg 5 (5.1%): Now that's intersting. These customers react well on BOGO offers (vtc rate 0.7), but seem to detest discounts (vtc rate 0.05).

- Seg 12 (8.4%): And here it's the other way round: Discounts are ok (vtc rate 0.7), viewed BOGO are completed to less than 1%.

Second group of segments - medium spending customers that have only been exposed to BOGO offers or to discounts only:

- Seg 3 (6.7%) and Seg 6 (9.9%): Have seen BOGO only. Seg 3 reacts ok with vtc rate of 0.7, Seg 6 considerably less with a vtc rate of 0.4.

- Seg 4 (6%) and Seg 8 (7.2%): Have seen Discounts only. Both Segments react ok with a vtc of 0.8 and 0.7 respectively. (One bigger difference here is that Segment 8 was exposed to a high share of informational offers, see remark below.)

3rd group of low spenders that could not be activated through offers:

- Seg 10 (8.7%): Quite regular customers (6 transactions a month vs. overall average of 8) that spend smell amounts only (total spending is only 25% of overall average). They view the offers on the app regularly, but do not react to BOGOs and discounts (vtc rate close to 0 for both).

- Seg 11 (5.7%): Similar to Seg 10 but have viewed no discounts. The average transaction amount is a little higher, so maybe there would be a chance to tickle them with some discounts. But probably they won't move to much.

- Seg 2 (3.8%): Similar to 11 but have viewed no BOGO. Maybe worth a try.

[Viz of Boxplots 2x vtc and total spending on 2nd y-axis]

[Viz of correlation from prop\_offertype\_amount to prop\_offertype\_viewed]

**4b) Conclusions on Demographic Groups**

In a next step I appended the segment predictions to the demographic customer information (note, I dropped all 2'175 customers for which we have no demographic info, so that left us with 14'167 customers to analyze because of the removed outliers. Should still be enough to receive valid results.)

The analysis focuses again on preference of specific user groups for BOGOs vs. discounts. In that respect the segments 5 (complete more than 2 3rds of viewed BOGOs, but no discounts) an 12 (just the opposite) seem the most promising to show a clear difference. I hoped for a clear pattern doing some scatterplots with different combinations of demographic features - unfortunately nothing was to be seen.

Looking at the features in isolation the pattern becomes clearer (and it is consistent for all pairs of segment comparisions, as can be seen in this notebook). Generally younger, lower income, male customers prefer discounts to BOGOs.

**5) Learnings**

Try to construct metrics / features that control for some of the variables.

Goal for starbucks

if possible substitute

- info for discount

- discount for bogo

It's more like an observational study than an experiment evaluation so this means

We typically cannot infer causality in an observational study due to our lack of control over the variables.

overall 65% cleaning & EDA (at least), 15% modelling, 20% analysis and report