Snapcommerce Analytics case study

Dhiren Kakkar

Glossary

- **LTR**: a customer's lifetime revenue
- **GMV**: a customer's gross merchandise value
- **CHECK_OUT_MONTH**: a customer's checkout month of their first booking

- **monthly_LTR** & **monthly_GMV**: above mentioned columns normalised by dividing the number of months since first booking made (to 1st january 2021)
- **Control group**: customers who did not receive a thank you call
- **Treatment group**: customers who received a thank you call

Breaking down the data

- 8287 out of 23759 customers received thank you call after their first booking.

- There are 26 unique months when the customers checked out from their first booking (2017/05 - 2019/06).

- 557 customers did not have their data recorded (missing data).

Recognizing temporal trends

- The temporal nature of the dataset influences the LTR, GMV & BOOKINGS since a customer with earlier CHECK_OUT_MONTH has a longer time window to make subsequent bookings.

- Hence, in order to statistically determine whether there is a significant difference between the LTR and/or GMV of customers who received the call vs of those who didn't, I **normalized** the columns by calculating monthly GMV and LTR (assuming an arbitrary present date as 2021/01/01).

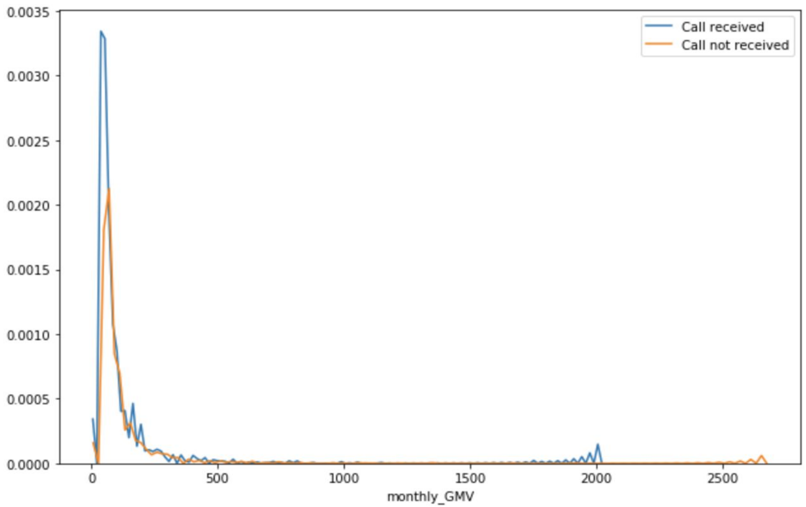
Filling in missing data

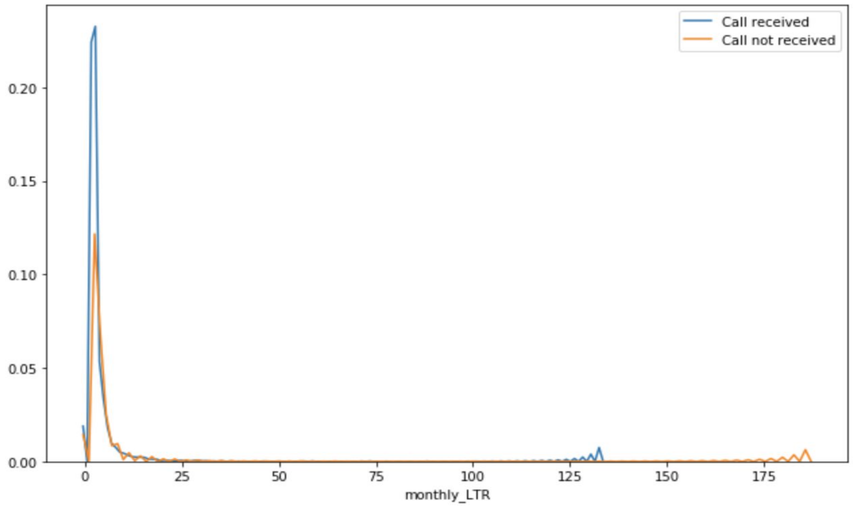
- Since a given customer’s LTR, GMV and BOOKINGS are a function of their CHECK_OUT_MONTH, the missing data for the 557 customers was filled in with the median column values for the given month.

- One CHECK_OUT_MONTH had two observations, both of which were missing. So, removed those two rows (can also fill as the mean of the preceding and subsequent month’s column values).

Target variable distributions

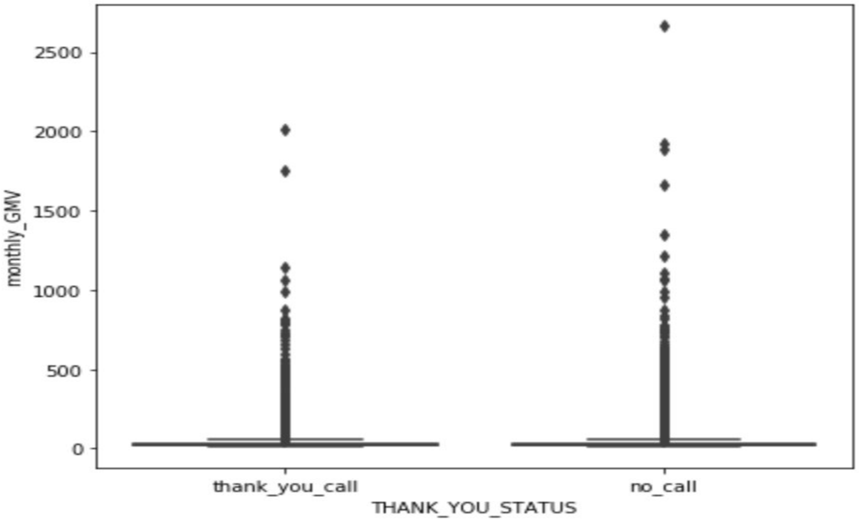
- The distributions of monthly LTR/GMV are **highly skewed** (and fails test for normality even with log transformation).

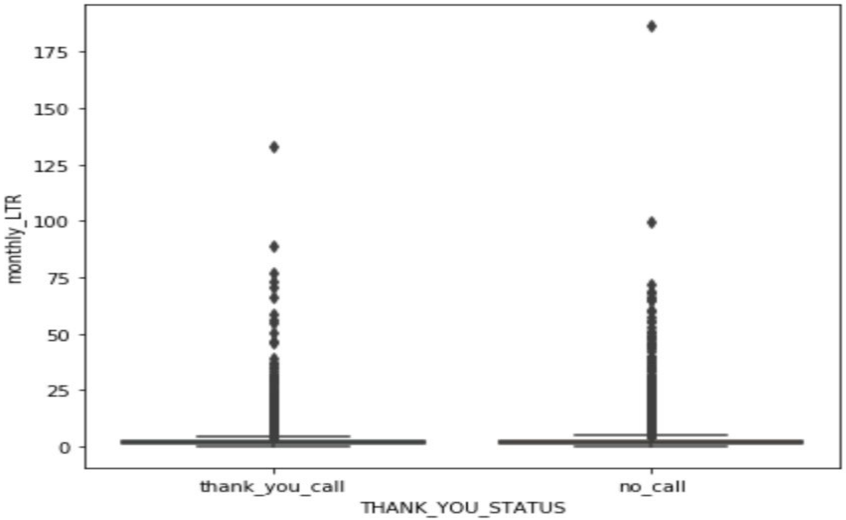




Analyzing impact of thank you calls

- No obvious visual differences between the two groups' monthly LTR/GMV.





Analyzing impact of thank you calls

- As mentioned before, the two groups are highly imbalanced. However, a general aggregation (median) for the two groups indicates that there **are no noticeable differences**.

- 2017/07 and 2017/06 have very large monthly LTR/GMV for the respective groups leading to the high skew. The group aggregations are still very similar without outlier observations.

```
Customers who received thank you call:
monthly_LTR              2.646282
monthly_BOOKINGS         0.063569
monthly_GMV             40.520384
dtype: float64

Customers who did not receive thank you call:
monthly_LTR              2.741980
monthly_BOOKINGS         0.072425
monthly_GMV             40.190457
dtype: float64
```

A/B testing: developing statistical hypothesis

General analysis can give an idea of how the treatment group differs from the control group. However, it doesn’t allow to conclude that this difference exists because of some causal relationship and not just by chance:

- **Two sample independent t-test** is often used to statistically determine whether two independent groups statistically differ (and not just by chance).

- For the purpose of this case analysis, consider a null hypothesis that there is no difference between the LTR and/or GMV of either group vs the alternative hypothesis that there exists a difference beyond just due to chance.

Potential issues in hypothesis testing and solutions

- **Imbalanced groups**: most A/B testing hypothesis testing methods assume that the control and treatment groups are same. However, unequal sizes of the two groups violates this assumption --- Resampling groups

- **Non normal distribution**: Since the distributions of monthly_LTR and monthly_GMV are highly skewed. General two sample t-test may not be the most powerful approach --- Non parametric testing

Hypothesis test results

- **Two sample independent t-test** (with and without resampling for label balancing) failed to reject the null hypothesis (at 70% power) that there is no difference between the two groups for either variable.

- Probabilistic non-parametric test, **Mann-Whitney U test** (allows to compare two similar distributions) rejects the null hypothesis suggesting that there is greater than 50% probability that a randomly drawn individual’s monthly LTR from the control group will exceed that of a treatment group.

Test conclusion drawbacks

- The non-parametric testing approach assumes the least assumptions which may have been violated by the dataset (especially since it was tested on resampled groups to ensure their distributions are similar).

- Despite paying special attention to ensure distributions/skew are similar for both groups, the test assumes identical distributions which **may not be the case**. However, this is the best approach to conclusively determine the A/B test for the given dataset (since even log transformations fails normality test).

Conclusions

- Both parametric and non-parametric statistical tests determine that sending a thank you call to customers after their first booking is **not yielding positive results in the future**.

- Even with conservative frequentist thresholds, it is statistically correct to conclude that the treatment group’s monthly LTR is not significantly different from that of the control group. The same applies for monthly GMV and monthly BOOKINGS.

Approach for additional experiments

- It might be useful to collect more treatment group samples as it improves the power of the A/B test (without explicitly generating more samples).
- While additional data outside of existing schema would not impact current testing metrics. Having further information regarding subsequent bookings details (booking location/duration of stay/cost/timestamps) can provide further context as to how customer patterns may vary between the two groups.
- This sequential information can be embedded in global features (and normalised using dynamic time warping approaches) to predict/simulate future involvements and empirically determine the efficacy of the experiment.

Part 2: Recommendation generation from the data

- There are multiple recommendation system approaches that can be employed to improve in-app experiences by understanding customer involvement patterns.

- This often requires pre-specified data fields/metrics to cluster customers by embedding information such as their sequential booking details, spending behaviour (other forms of involvements like browsing behaviour) in the form of global and local features to target similar groups together.

The workflow

While the process may require frequent back and forth to ensure business goals are met, the development lifecycle can be broken down into two main parts:

- **Data collection/aggregation**: As mentioned, an important step in the process is to analyze and identify features that allow accurate grouping of users based on their booking patterns. (more on the dimensions and aggregations later)

- **Iterative testing/development**: Once a reasonable baseline model is implemented, iteratively testing on a subset of users can often help identify and generalize the impact of recommendations on their behaviour. This approach often employs A/B testing on a few app/web interaction events.

Improving user experience using collaborative filtering

- Collaborative filtering approach is a method to predict for a user by understanding relations, instead of defining ratings for each dimension. This approach pools the information of similar users and can provide recommendations by learning from both local and global features to group similar users together.

- Directing recommendations by understanding the sequential booking data of a given customer against that of other customers' with similar sequential booking data can have great predictive power to recommend hotel room packages that may (statistically) entice them more.

- Since users may have differing/unevenly spaced sequences, methods like **dynamic time wraping** can be used to identify patterns in varying time series windows.

Quantifying data as actionable insights/metrics

- **Sequential insights**:
    - Since traditional collaborative filtering lose the sequential information about a user, global & local features can be engineered in order to more accurately pool similar customers together. This often involves embedding series data in the dimensions (features).
    - There are a number of features which can capture the sequential nature of the data without having a traditional time series structure (discussed in next slide)

- **Browsing patterns (More useful for A/B testing)**:
    - CTR (clicking through recommendation)
    - CR (conversion rate): bookings made for the given recommendation

Potential insights (KPIs) from data aggregations

- Booking details (country, hotel, duration, cost, booking date, total members):
    - Total bookings made so far: travel frequency
    - Time gap between bookings: sequential information about travel frequency
    - Cost change between bookings normalised by total members: financial capacity indicator
    - Hotel star rating: vague financial capacity indicator
    - Change in hotel star rating between bookings: sequential financial capacity change indicator
    - Hotel destination: vague financial capacity indicator
    - Previous destination: sequential travel details
    - Duration of stay: travel details

*Most of these local features can be compared against their global aggregation values. This allows to recommend products that similar users are engaging in:

- Booking seasonality: booking behaviour of similar users at a given time window

Understanding the impact of the recommendation system

- Early stages of recommendation systems often require deliberate placement of recommendations (both content and UI) for only a subset of the population in order to carry out A/B testing. These approaches help in determining the efficacy and relevance of the predicted recommendations.

- Most hypothesis testing methods often focus on click-through rates (CTR) and conversion rates (CR) of the treatment and control groups. This not only allows to quantify the interactions with the recommendations, but also carry out traditional A/B testing approaches to statistically determine this impact.