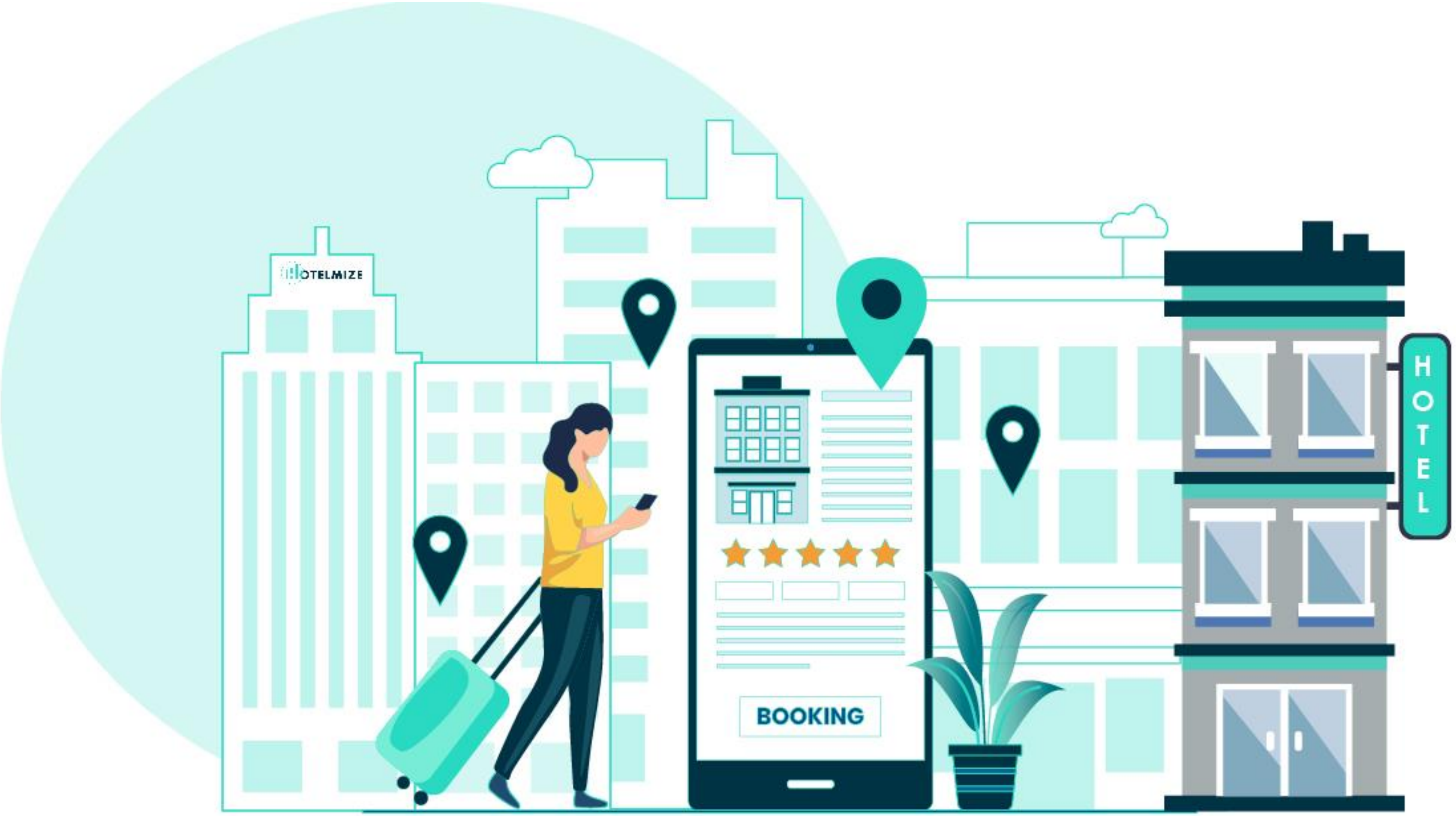
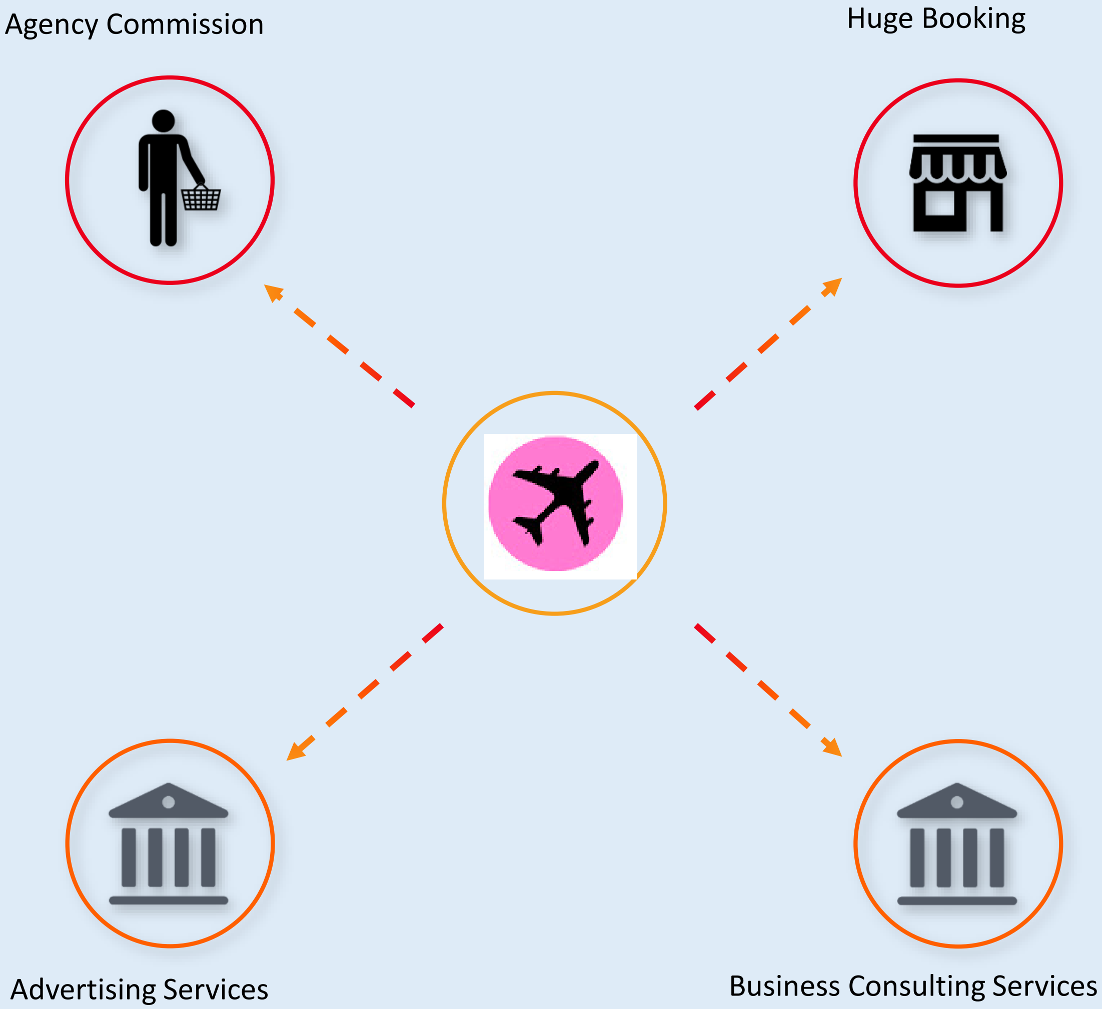
Mastercard Case Study



Satyam Neelmani



To provide end-to-end Data-Driven solution for a major Online Travel Agent in India

- Problem Statement

Business Model of Online Travel Agency

1. Huge Booking: Online Travel Agency books flights, hotels in bulk at discounted rate from service providers and sell to customers in the form of holiday packages.
2. Agency Commission: Online Travel Agency earns from the commission they get whenever there is any booking for cab or flights by customer from their platform.
3. Advertising Services: Online Travel Agency provides a paid platform for hotels so they get recommended to customers.
4. Business Consulting Services: A lot of data of customers is collected which an Online Travel Agency uses to consult other complement players to drive their business growth.

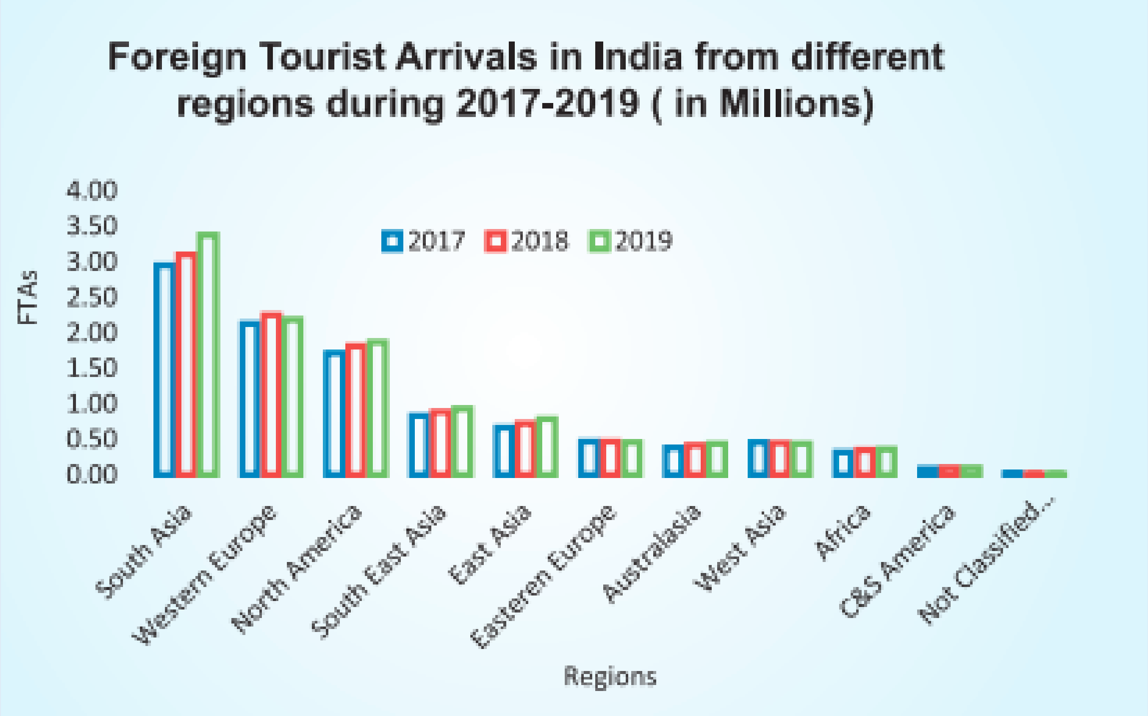
Case Assumptions

Timeline since 2020 has been an uncertain situation, so the data of 2019 has been used in this case study and also in the empirical modelling.

Data Sourcing

1. To identify customer intent to travel to India, online survey at social media, personal mails can be conducted. It will collect attributes like covid-19 facilities/protocols customers expect, amount they are willing to pay, mode of transaction, purpose of tourism, etc.
2. Transaction amount, transaction number, transaction date and time, transaction type and domestic travel details, **Covid cases status, vaccination status, spending capacity, number of days of visits** will be used as important features to segment customers. We will include Recency, Frequency and Travel Priority, Airport demography, etc.

Best Inbound Markets



Reference

Go To Market Strategy

Bargaining Power of Supplier : Moderate

Less number of COVID protocol compliant players

Bargaining Power of Consumers : High

Few people will agree to come initially

Threat of Substitutes: Low

Very few players with Liquidity enough to restart business

Threat of New Entrant : Moderate

Competitor Online Travel Agency are expanding its base

Understanding segments that are probable to visit India

Business Travellers

Young Tourist Travellers

Academicians & Family

Medical Travellers

Customers' Value

High value customers

Mid value customers

Low value customers

Considerations

High value customers
Minimum variation within
Homogenous segment
(Distinctively defined)

Max. variation between
Heterogeneity

Mid value customers
Minimum variation within
Homogenous segment
(Not distinctively defined)

Max. variation between
Heterogeneity

Low value customers
Minimum variation within
Homogenous segment
(Distinctively defined)

Empirical problem definition

Unsupervised Learning | K-Means Clustering
Identify homogenous cluster(s) of high value customers on the basis of given feature values

n = Total no. of observations
k = No. of clusters
P = No. of features
C_k = set of k^{th} cluster

Properties

1. $C_1 \cup C_2 \cup \ldots \ldots \cup C_k = \{1,2,3 \ldots\ldots, n\}$
   Each observation belongs to atleast one of the K-Clusters; Collectively exhaustive
2. $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$
   The clusters are non-overlapping; Mutually exclusive

Objective function | Minimize within cluster variation

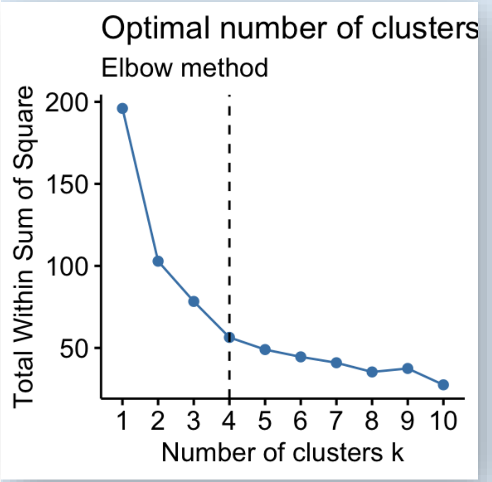
$$\underset{C_1, C_2, \ldots, C_k}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in Ck} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

K-Means Algorithm

1. Randomly assign a number from 1 to K, to each of the observations. These serve as initial cluster assignments for the observation
2. Iterate until the cluster assignments stop changing
   i. For each of the K clusters, compute the cluster centroid. The k^{th} cluster centroid is the vector of the 'p' feature means for the observations in the k^{th} cluster.
   ii. Assign each observation to the cluster whose centroid is closest (by Euclidean distance)

Limitation

- K-Means method aims to find the local minima (by limitation of the optimization problem)
- Since it finds a local minima, the result/outcome is dependent on the initial cluster assignment of each observation (step 1). Hence, for a better outcome/solution, we should iterate K-Means method for 1000 or 10000 times, to get a cluster with the minimum variation within clusters. This can be a proxy for global minima optimized solution.



Model Optimization

Determine optimal K (no. of clusters) for K-Means

1. Elbow Method
   Calculate the Within-Cluster-Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS becomes first starts to diminish.
   In the plot of WSS-versus-k, this is visible as an elbow.

2. Silhouette Method
   The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation). The range of the Silhouette value is between +1 and -1. It is desirable to have a high value, i.e value ~1.

$$s(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}$$

Where,

$$b(i) = \min_{i \neq j} \left\{ \frac{1}{|C_j|} \sum_{j \in Cj} d(i,j) \right\}$$

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in Ci\, i \neq j} d(i,j)$$

s(i) : Silhoutte value for each data point ‘i’
a(i) : Measure of similarity of ‘i’ to its own cluster
b(i) : Measure of dissimilarity if ‘i’ from points in other cluster
d(I,j) : Euclidean distance between ‘i’ and ‘j’

Implementation

Identify high value segment

If K= 3;
The cluster whose mean value of transaction amount as a measure of spending capacity is highest of the 3 clusters; The “High value segment”, whereas the other two being mid and low value segments.
If K = 2; High and Low value segments
If K > 3; Define high value segment subjectively by looking at the feature values and higher mean transaction amount of the cluster

Best feature selection

p = No. of features
i.e, p = {Types of data collected; Ref. Slide 3}

To select best features that identify the high value segment distinctly, calculate and iterate to minimize within-cluster-sum of squared errors (WSS) using forward or backward propagation method, analogous to linear regression

Thank You