Customer Segmentation and Price Optimization Strategies for Singi's Kitchen

A Mid-term report for the BDM capstone Project

Submitted by

Name: Sharad Anirudh Jonnalagadda
Roll Number: 23f2000690
Email id: 23f2000690@ds.study.iitm.ac.in

IITM Online BS Degree Program,

Indian Institute of Technology,

Madras, Chennai Tamil Nadu, India,

600036

Contents:

1. Executive Summary

Singi's Kitchen is a community restaurant that serves both vegetarian and non-vegetarian delicacies at affordable rates and has both Dine-in and takeaway options available, the restaurant opened its doors in the September of 2024. The restaurant faces many problems including lack of proper inventory database, lack of knowledge regarding their customer base and proper menu optimization strategies, which are undermining the potential of this restaurant. This project was undertaken to help Singi's kitchen develop more revenue, to understand the customer demographics in the community of Saket Bhu: Sattva and to bring effective changes in the menu to increase their profits.

Based on the problem statements provided in the proposal, which primarily works on, "Customer Segmentation and targeting", "Menu Items optimization" and "Inventory Management", The established goals for the mid-term are as follows:

- To develop a model named ***potential rating score,*** which would predict if a household would continue ordering in the future based on multiple features which includes frequency of orders, age, cultural background and etc.
- To develop a ***grid heat map*** which would color-encode every villa in the community based on their potential rating score, ranging from red (low potential score) to blue (high potential score), while using gradient of the two for intermediate values.
- To construct additional heat maps to analyze parameters such as average daily spending, average daily profits and average potential scores for each phase of the community.

2. Proof of Originality of the data

(1) Photos regarding the owner and the organization:

Image 2.1: Meeting with the owner Mr. Rama Raju Singi

Image 2.2: Front view of the restaurant

(2) Letter of Authorization from the Organization:

Letter of authorization by Singi's Kitchen

Date: 2nd November, 2024

To whom it may concern,

Singi's kitchen authorizes Mr. Sharad Anirudh Jonnalagadda to work on his project "customer segmentation and menu optimization techniques". We agree to share all relevant data for the project, which includes but not limited to, sales data, inventory data, bills and slips. We also acknowledge that all of the details taken for building the project is only for educational purposes. We wish all the best to Mr. Sharad Anirudh Jonnalagadda and we are eagerly looking forward on seeing the insights that will be shared by him at the end of the project and ultimately, to work upon the areas which need improvement.

Sincerely,

Mr. Rama Raju Singi (Owner of Singi's Kitchen)

Contact: +91 8106384666

(3) Recorded Video with the manager of the Organization:

YouTube link: https://www.youtube.com/watch?v=IkaDFap2bz8

G-Drive link: https://drive.google.com/file/d/1--63aWQmt5J__WR9cuwN1oH7IIoBUNcy/view?usp=drive_link

3. **Metadata**:

The dataset consists of 269 rows, covering the entirety of the community, each of these records have been covered based on certain features which has a pivotal role in constructing the *grid heat map*. The features (or the columns of the dataset) include; "average frequency of orders daily (x_1)", "current food rating (x_2)", "previous food rating (x_3)", "average age of the household (x_4)", "cultural background (x_5)" and "Dietary preferences (x_6)". Then the next immediate column of the dataset is the potential score, which has been calculated by using potential rating score formula which would be discussed in later sections. There are few other columns in the dataset, but they are not essential features for the potential formula, rather they are *derived features*, which have been collected to understand more about each of the household in the community.

Data sources with regard to sales, which includes sales of each individual villas, daily profit made based on the orders, average frequency of orders per day for each household was collected from the sales data provided by the restaurant and data collected from the takeaway orders which are made Via WhatsApp. The restaurant didn't provide with the data regarding takeaway orders, I had collected the takeaway orders of each villa every day and stored it in the dataset. The data collection commenced on 3rd of October and was collected up until 3rd of November. The data collected for each villa was then compiled into a grand dataset with 10 columns and 269 rows as mentioned above

For the data cleaning part, standard procedures were followed, which includes, standardizing villa numbers converting the villa numbers from alphanumeric quantities to numbers, ranging from 1 to 269, not considering dishes which made very minimal impact to the profits of the restaurant, not considering dishes which were seasonal and considering twelve most important dishes out of a menu of 45 items for a more robust analysis. For customer side of things, a google form survey was created and circulated throughout the WhatsApp groups, a total of 262 responses were received in the end, for the remaining responses, I had physically visited these villas to collect the necessary information. Despite the efforts, 4 households were unavailable and didn't provide with the necessary information, for these villas, sales were observed and the columns were filled based on their sales data. Links have been provided for both, the finalized dataset as well as the google form survey:

Link for the finalized dataset:
https://docs.google.com/spreadsheets/d/10AHnCaQo9BbXDrfpOif6NTgj4mov6208/edit?usp=sharing&ouid=104943617588248477463&rtpof=true&sd=true

4. **Descriptive Statistics:**

The dataset was analyzed to understand its central tendencies and variability involving key features, all of the analysis was done in MS EXCEL by using in-built formulae for interpreting all of the standard values such as Standard deviation, Mean, 1st Quartile,2nd quartile, 3rd quartile and etc. below are the key statistical highlights:

- **Frequency of orders** (x_1): The average frequency of orders per household was found to be 0.55 orders per day, with a standard deviation of 0.26. The highest recorded frequency was 1.0, while the lowest was 0.11. The key takeaway here is, there is a significant variability in frequency of orders in the community.
- **Current Food Ratings** (x_2): The average customer rating for food was 3.07 on a scale of 1 to 5, with a standard deviation of 1.41. A considerable portion of the community has rated above 4, which indicates a positive perception for the restaurant, but on the flipside, there are quite a few low ratings as well which were as low as 1.
- **Previous Food Ratings** (x_3): Historical ratings averaged slightly lower at 2.95, with a standard deviation of 1.15. The key takeaway here is, the food ratings have considerably improved over the course of time, which indicates a positive shift for the restaurant.
- **Average Age of Residents** (x_4): The mean age across all the households was approximately 39.5 years, with a standard deviation of 12.5 years. The youngest household in the community had an average age of 18 years, whereas the oldest household in the community had an average age of nearly 60 years. The key takeaway here is, the average customer base of the restaurant is middle aged people, which is an important factor to consider for the restaurant.
- **Potential Scores:** The calculated potential scores ranged from 20.1 to 80.4, with an average of 51.1. Scores below 35 were categorized as low potential, while scores above 75 are categorized as high potential households**.**
- **Daily Spending:** on an average households spend nearly ₹96.23 per day on orders, with a standard deviation of ₹74.14. The highest spending done by a household in the stipulated period of time was ₹346.50, while the lowest recorded spending by a household was ₹4.80.
- **Daily Profit**: The restaurant generated a daily profit of nearly ₹39.31 per household in the given stipulated period of time, with a standard deviation of ₹29.34. Profits recorded range from ₹2.40 all the way up until ₹138.60 on an average on a particular given household. The key takeaway here is, there is a high variability in profits generated and most importantly each household is performing differently in terms of profits.
- **Graphs and charts**: A violin chart of daily spending by a household on the most ordered items and a tabulated graph representing central tendencies of each feature in the dataset have been created manually with the help of online software including Lucidspark and Canva, both of which are provided in results/Findings section. Both of these are extremely insightful and important.

5. Detailed explanation of the Analysis Process/Method:

The explanation of methods employed will be divided into three main categories, which are, "Data collection and Data cleaning", "Derivation of the potential formula" and "Construction of the grid heat maps", the detailed explanation is as follows:

Data collection and Data cleaning:

The analysis began by compiling 269 rows in the dataset which represents each villa in the community. The sales data was gathered through multiple sources including sales data provided by the restaurant regarding all the customers who visited the restaurant from 3rd of October to 3rd of November and similarly, for all the orders placed via WhatsApp, the sales data was collected manually starting from October 3rd to November 3rd. Also, a google form was constructed, keeping in mind that the questions asked were apt and concise and also, they are framed in such a way that its easy to interpret the answers and apply those in the potential rating formula.

The process applied in data collection and cleaning have been mentioned below in a very apt and concise manner:

- Villa numbers were standardized, converting alphanumeric identifiers into sequential numbers from 1 to 269.
- Dishes with minimal sales impact or those deemed seasonal were excluded to focus on the most influential items.
- Features such as Cultural background, dietary preferences and age group were extracted from survey responses, while the missing data from four households was estimated based on the sales trends of those particular households.
- Then the dataset was properly organized which consists of 10 columns out of which 6 are important key features in the potential rating formula and others are derived features which are essential for additional insights of the community.
- The process of data collection was mainly collected on paper for over a month and in the end, all of the necessary data was fed into the finalized dataset.
- Questions asked in the google form survey were simple questions which asked the customers about their cultural background, their current rating of the food, their previous rating of the food, average age of the household and dietary preference of the household. This survey is extremely important as it directly correlates to the performance of the potential rating formula. (A link for the google form survey is given below):

Link for the G-form:
https://docs.google.com/forms/d/e/1FAIpQLSew3mM4RHV6ahk_ZhgJyA6g992KVYTfUyyBQFPbLtL6NPWAyA/viewform?vc=0&c=0&w=1&flr=0&fbzx=3254787555208146908

Derivation of the Potential Formula:

The potential rating score formula was developed to predict customer loyalty and likelihood of future engagement. This model integrates six key features, each assigned a weight based

on its importance to customer behavior, given below is the developed formula:

$$r(x_1, x_2, x_3, x_4, x_5, x_6) = 30x_1 + 6L(x_2, x_3) + 2\left(10 - \frac{x_4}{10}\right) + N(x_5) + M(x_6)$$

Where:
x_1: Frequency of Orders
x_2: Current food rating
x_3: Previous food rating
x_4: Average age of residents
x_5: Cultural background (North Indian, South Indian, Other)
x_6: Food preferences (Veg, Non-Veg, Both)
r: potential score of that household
L(x_2, x_3) is defined as:

$$L(x_2, x_3) = \frac{x_2 + x_3}{2}, if \quad x_2 - x_3 \geq -1$$
$$= 0 \quad , if \quad x_2 - x_3 < -1$$

N(x_5) is defined as:

$$N(North\ Indian) = 6$$
$$N(South\ Indian) = 5$$
$$N(Other) = 2$$

M(x_6) is defined as:

$$M(Veg) = 3$$
$$M(Non - Veg) = 5$$
$$M(Both) = 7$$

Explanation of the formula and how it was derived:

Overall premise of the formula: Firstly, the whole formula has been developed in such way that the maximum possible rating is 100(theoretically) and the minimum possible weight is 0 (theoretically).

Difference of weights to different parameters: Different weights have been allotted to different parameters based on importance they have on determining customer's loyalty in the future.

Feature 1 (x_1) order frequency: According to research article published by "Monetha" (link:https://business.monetha.io/blog/customer-loyalty/purchase-frequency/), Customer loyalty strongly depends on the frequency of orders they place, which is also an important observation found from the data collection process in this case. Therefore, considering its importance a good portion of the weight, which in this case is 30% had been allotted to the feature 1 in the formula.

Feature 2 & 3 (x_2, x_3) Heuristic function: According to a book published by springer (Customer satisfaction, loyalty behaviours, and firm financial performance, link: https://link.springer.com/article/10.1007/s11002-023-09671-w), Higher satisfaction levels over a course of time in any given firm or organization, will improve customer loyalty drastically. Also, in data collection of over a month a similar trend was observed. Hence, a simple heuristic function was developed where two features, previous food rating and current food ratings are considered and if the subtraction of those two is less than 1 (which implies customer is growing very unsatisfied) the value of the function is nil, otherwise average of the two is considered. As it is also an important contributor to customer loyalty, 30% of the total weight was assigned to it. (Total weight here is calculated by considering $x_2 = 5$ and $x_3 = 5$, ultimately, after doing the calculations, the resultant total weight is 30%)

Feature 4 (x_4) Average age feature: It was observed after thoroughly going through the dataset that if the average age of a household is high then daily money spent on the hotel will be considerably low. This proposition is backed up by the research article of published by the

journal of business economics (link: https://link.springer.com/article/10.1007/s11573-016-0834-4) which states, as people age they become more selective and less inclined on staying loyal to a single business. Considering this, a simple function involving average feature has been developed where the age will be divided by 10 and then subtracted by 10, and ultimately will be multiplied with the allotted weight for the given feature. The reason for using a simple appropriate function is to ensure that by growing age of the household the less is the resultant rating obtained, developing an essential inverse proportionality between the potential rating score and the age feature. The weight has been 2 which implies, the overall weight of this feature is 20%.

Feature 5 & 6 (x_5, x_6) Factor functions: Another interesting detail that was found after going through the dataset was, customer's cultural background and customer's dietary preferences also influenced if they would stay loyal to the restaurant in the future or not. To mention all of the nuances in a very simple manner, it was noticed from the dataset, 6 in every 10 north Indian customers stayed loyal, 5 in every 10 south Indian customers stayed loyal and 2 in every 10 "other" customers stayed loyal in the given stipulated period of time of one month. Similarly, 3 in every 10 purely vegetarian customers stayed loyal, 5 in every 10 non-vegetarian customers stayed loyal and 7 in every 10 customers who preferred both type of dishes stayed loyal in the given stipulated period of time of one month. By this, it could be understood that although these are not major contributing features, they make an impact on customer retention and loyalty too. The original intention was to develop both of these functions in such a way that each contribute 10% to the total rating score, but practically its not possible to assign a rating score of 10 to any of the customers as the sales data justifies, no such cultural groups or dietary groups were entirely loyal as to give any customer a score of 10.Hence practically the highest a customer can get in this regard is 6 with respect to $N(x_5)$ and 7 with respect to $M(x_6)$, This ensures that the formulation is practically viable and follows the implications set by the dataset, which in turn makes the formulation more practical and robust. Although these features are minor, they add depth to the formulation and also ensuring no house gets a perfect score of 100 is another indicator that the formulation is practical and robust.

Construction of grid heat maps:

To visualize the grid heat map, a satellite image of Saket Bhu: Sattva was captured using google maps. A real time image was captured using google maps satellite view, keeping in mind that the quality of the picture should be high, such that each and every villa in the community are clearly visible. Then each villa was manually numbered from 1 to 269, by using the Canva software. After that the following work was done:

- **Color encoding:** each and every villa in the community was color-encoded, where the colors ranged from blue (low potential rating <25) and red (high potential rating >75), where gradient of the two colors were used for any intermediate quantities. The gradient of two colors based on the rating score was achieved using Canva software, where potential scores were inputted into the color palette to assign appropriate shades to each unit.
- **Construction of phasal maps:** after successful completion of grid heat map, phasal heat maps were developed, which included dividing the community into different phases, and then color-encoding these phases based on the average potential of the phase, average daily profits generated due to that phase, average money spent daily by that phase. The division of community into phases was done based on their building dates, which resulted in formations of rectangular blocks which reveal deep insights about the community. The phases are significantly important, as it would allow the

restaurant to develop phase-based strategies, which would optimize their profits and revenue without focusing on each and every villa, generating a high-level overview of the community's customer base. The process of color-encoding is the same as mentioned above, using Canva software, but specifically for this case, larger blocks were created to vividly represent the phases of the community.

The constructed heat maps have been provided below in the results/findings section.

6. Results/Findings:

By the end of mid-term, the following are all of the results/findings:

- **Potential formula:** A mathematical formula was developed to calculate the potential rating score for each household. The formula incorporates six features: frequency of orders (x_1), current food rating (x_2), previous food rating (x_3), average age of the household (x_4), cultural background (x_5), and dietary preferences (x_6). The final formula is given as:

$$r(x_1, x_2, x_3, x_4, x_5, x_6) = 30x_1 + 6L(x_2, x_3) + 2\left(10 - \frac{x_4}{10}\right) + N(x_5) + M(x_6)$$

$L(x_2, x_3)$ is defined as:

$$L(x_2, x_3) = \frac{x_2 + x_3}{2}, if \quad x_2 - x_3 \geq -1$$
$$= 0 \quad , if \quad x_2 - x_3 < -1$$

$N(x_5)$ is defined as:

$$N(North\ Indian) = 6$$
$$N(South\ Indian) = 5$$
$$N(Other) = 2$$

$M(x_6)$ is defined as:

$$M(Veg) = 3$$
$$M(Non - Veg) = 5$$
$$M(Both) = 7$$

- **Finalized dataset**: The finalized dataset contains 10 columns and 269 rows, capturing customer data, sales information, and calculated metrics.
  Link for the finalized dataset:
  https://docs.google.com/spreadsheets/d/10AHnCaQo9BbXDrfpOif6NTgj4mov6208/edit?usp=sharing&ouid=104943617588248477463&rtpof=true&sd=true

- **Grid heat maps:** After successfully constructing the potential rating score formula, grid heat maps were constructed over a captured satellite image, where each specified unit has been color-encoded from blue (<25 potential rating) to red (>75 potential rating), and using gradient colors for intermediate values. Grid heat maps for individual villas and also phasal heat maps are provided below:

Image6.1: Potential rating scores of Saket Bhu: Sattva

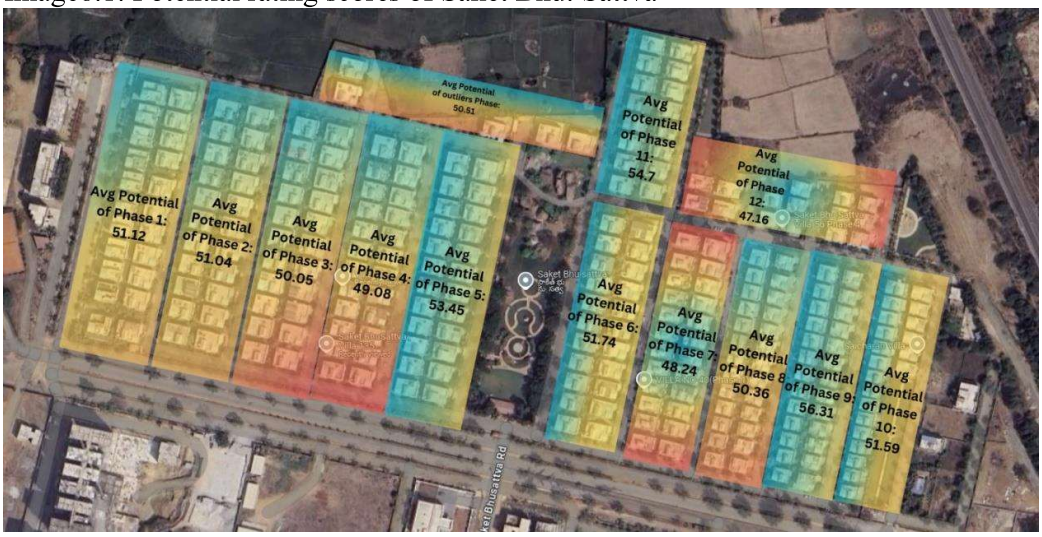

Image6.2: Phasal Grid heat map based on potential rating scores

Image 6.3: Phasal grid heat map based on average money spent by a household on restaurant on an average in a day.

Image 6.4: Phasal grid heat map based on average profits generated by the restaurant on a household in a phase at an average.

- **Descriptive Statistics:** The dataset was thoroughly analyzed and all of the necessary standard values and central tendencies of all the features in the dataset were calculated, part that follows contains a tabulated form of the central tendencies and a violin chart, which showcases the distribution of each important ordered item with respect to daily spending.

	Statistics	x1	x2	x3	x4	Potential	Daily Spending	Daily Profit
0	count	293.0	293.0	293.0	293.0	293.0	293.0	293.0
1	mean	0.554266	3.068259	2.94744	39.549488	51.123891	96.228328	39.311433
2	std	0.260801	1.407702	1.155438	12.543422	11.924657	74.142579	29.345066
3	min	0.11	1.0	1.0	18.0	20.1	4.8	2.4
4	25%	0.32	2.0	1.9	28.0	43.5	37.2	16.5
5	50%	0.56	3.0	3.0	40.0	51.7	75.0	30.6
6	75%	0.75	4.0	4.0	51.0	59.1	136.8	54.6
7	max	1.0	5.0	5.0	60.0	80.4	346.5	138.6

Image 6.5: Central tendencies for each feature in the dataset is calculated and tabulated.

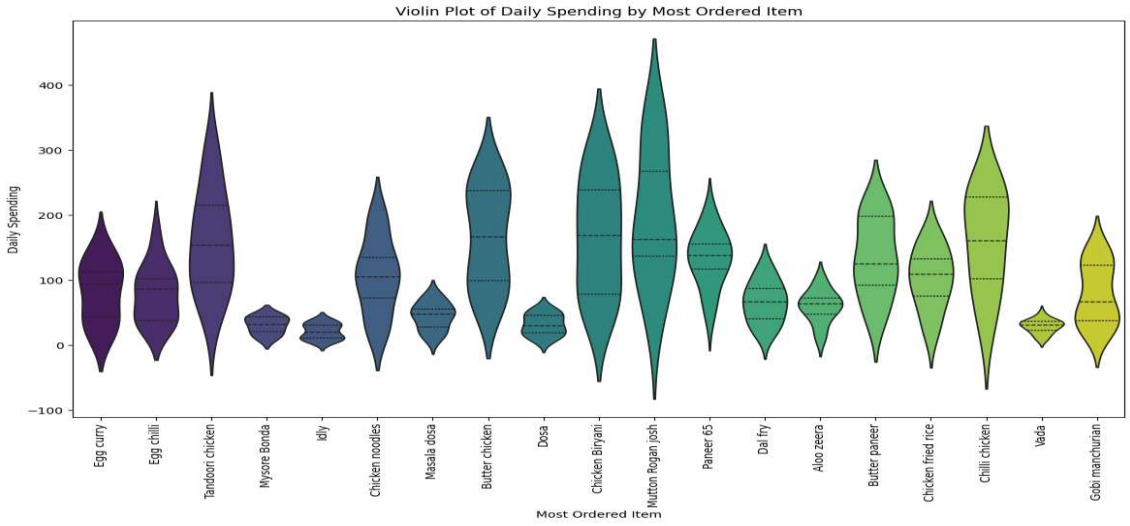

Image 6.6: A violin chart that is very useful to understand quartiles of the distribution