Tactics to Maximize the Lifetime Value Potential

December 7 2022

By
Bengal880: Lilly Pan, Cathy Tang, Helen Zhu

For
Habib Ghazi Talent & People Development People and Organization, Grameenphone
Syed Shakil Ahmed Commercial Planning & Development Commercial, Grameenphone
A. Z. M. Rased Commercial Planning & Development Commercial, Grameenphone
Syed Masud Mahmood Talent & People Development People and Organization, Grameenphone

Executive Summary

Grameenphone is our sponsor who cares about picking the optimal set of actions that maximize the customer lifetime value and what actions should be taken. We accomplished the primary goals by utilizing the Gamma-Gamma model, Customer Lifetime Value(CLV), K-means, logistic regression, and correlation to create customer profiles based on different coefficients we manually selected from the large dataset. We chose to use the process of CLV because it implicates how they impact a firm's profit and how actions affected the overall profits. Our results indicated that the most valuable customer groups are more likely to have higher purchasing frequency, recency, and high lifetime values. We also found the correlation between total revenue and other variables. The results showed that customers in the third group heavily rely on Data Revenue, Charged Data Usage, and Outgoing Minute Usage. It gave insights about consumer behaviors by groups, what factors have the greatest influences, and increases service quality.

For the importance of our outcome from the business perspective, the model had the potential to calculate customer lifetime values and future predictions by classifications. In this way, the company can make suitable marketing investments toward its clients. However, we also had some concerns and limitations in selecting variables. Due to a large number of features, we only selected 11 factors related to total revenue. Therefore, it may have some biases. Although our team can not solve the difficulty of our building model, it is also considered as our future plan to increase accuracy and comprehensiveness.

1 Problem Statement

Our sponsor, Grameenphone, is a leading telecom service provider in Bangladesh. The problem statements are optimizing and increasing customers' lifetime values, understanding what factors affect customers' decisions and creating a predictive model with revenue as outcome variables for future prediction.

Our team used tactics of calculating customer lifetime values, k-means, correlation, classification, logistic regression, and other statistical methods to answer research questions and had deliverable messages. In our project, we needed to address many problems, including data cleaning with various categories in a larger dataset and model the customer lifetime value by its frequency. We aimed not only to build relationships with customers but also to allow our sponsors to know the right customers that typically generate the most profit for the company. The highest priority of the objective was to find out how to increase customer satisfaction and what factors caused these customers to become the most valuable people to the company based on our results. People who care about our results are our clients, especially for marketing divisions and their customers. Therefore, both producers and consumers benefit from our outcome. By evaluating our models to identify the most influential variables that motivate customers to spend more, Grameenphone can improve their financial benefit as a business and consequently can devote more attention to optimizing plans and services. On the other hand, customers can ensure that they have made suitable marketing investments by looking at which cluster they are more applicable to join. Scientifically, our model applied several statistical approaches to calculate the ratios and similar characteristics by grouping the data into four parts. It is easy to make reliable results and help for further study.

2 Ethical Considerations

Our envisioned purpose is to form a win-win situation where users can enjoy the service that best suits them, and our client companies can increase profits. Suppose malicious people obtain our model, they may use our prediction results to maliciously exploit users, such as over-marketing them to suit their preferences and attracting customers to buy services or products. However, the service or product is of no use to the users at all. It will only make them spend more money. We expect to provide our client company with the best means of serving customers without over-consumption so that customers are willing to use the company's services and products. If a company entices customers to buy services and products that do not fit them by digging into their habits and interests, one day, they will finally find out and give up their cooperation with the company. Implementing this "trick" is unethical and will only increase the customer churn rate and reduce the lifetime. In the long run, the short-term profit is smaller than the long-term profit. In order to prevent this situation, the assurance we can take is to communicate our concerns and considerations with our client. We want to make sure that they will ensure that they will keep their company's network secure so that our models will not be stolen by malicious people or companies.

3 Literature Review

Our team is not the first and only group to focus on customer lifetime value. Therefore, as our team brainstormed to solve our research problem, it was inevitable that we would need to consult the technology and experience of other people. Accordingly, we consulted the following three pieces of literature to help identify the algorithms and models we potentially use.

3.1 Modeling customer lifetime value in the telecom industry.

Petter Flordal and Joakim Friberg looked at what drivers behind customer lifetime value and how they impact a firm's profit. Based on probability and statistical theory applied in a practical marketing setting, they aim to model the customer lifetime value by Markov chain modeling and applied ordered probit regression to analyze the survey data. In addition, they described the dynamic relationship between a customer's preferences and the profit it generates during its lifetime. This paper also inspires our thinking by mentioning similar background knowledge about our project: customer lifetime value is associated with the time customer retains. Based on this article, our team gained a deeper understanding of our research question and expanded our thinking[1].

3.2 Teleco customers churn prediction using machine learning

Abdul Razak and Hazim Wahid present customer churn prediction based on the usage pattern using a machine learning prediction model, namely linear regression, random forest, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and decision tree. Also, they compared these algorithms through accuracy, recall, precision, F1-score, and area under the curve (AUC). Our team will only think about some of the models they mentioned, but their application and evaluation of these models is a good reference for our team to select a method[2].

3.3 Customer purchase prediction through machine learning

Hannah Sophia Seippel provides a real-life example to analyze customer behavior and find purchase trends. He compared models further and gave insight into the performance differences of the models on sequential clickstream and the static customer data by conducting a descriptive data analysis and separately training the models on the different datasets. It is similar to what our team will do for our project and gives our team references. Our team can get some experience from it and refer to its methods and processes[3].

4 Project Criteria

Based on our data and the purpose of our project, we list the standard criteria that guide us on the choice of algorithm. Here are the main criteria for our team.

4.1 Criteria for (1) Data Cleaning

- **Data Validation:** The program needs to provide relevant, useful information. All qualitative and quantitative categories should be represented in our data.
- **Completeness**: Data should not be left blank or missing. We can throw away the rows or columns containing missing data or input the missing data(such as 0 or the mean)
- **Type of Data:** Covert the same type of data to the same unit of measure.
- **Speed:** No amount of data can be processed in more than 15 minutes

4.2 Criteria for (2) Math Modeling

- **Languages:** We should be using Python or R to generate our projects.

- **Tools/ Strategies:** Pycharm, VS code or Rstudio
- **Scaling:** The scale ranges from 1 to 0, with 1 representing the ideal score for identifying the model that fits the database the best.
- **Accuracy:** The models should have fewer errors and mitigate risks to make final decisions and optimize actions. The error between the model's predictions and the real data should be within 5%.
- **Speed:** The time for training data cannot be more than 15 minutes.

5 Selected Solutions

5.1 Data cleaning

We used the approach of removing records with missing values for data cleaning. Removing records is a fundamental method for deleting unwanted columns or rows. We need to look at revenue and how the client's users relate to these factors. Our research shows inactive users do not generate revenue, which is useless and needs to be cleaned up. Our team wanted to delete data that was meaningless and irrelevant to revenue, improving the time and accuracy of the model we built later. Based on the language we use, Python, we use NumPy and Pandas packages to process the data.

In addition, since our sponsor gave us a large amount of data, we decided to split the data to reduce the amount of computation and iterations. We chose the Gamma-Gamma model to evaluate the correlation between average purchase amount and purchase frequency and calculate the possible future sales and predicted lifetime value for each customer. Based on the customer lifetime value, we split the data into three parts. In this solution, we used the GammaGammaFitter package in python to build the GammaGamma model.

5.2 Math modeling

We need to analyze our clients' users to determine their preferences for our client's products. We divided our clients' customers into different levels through revenue and find out what the users of each level have in common. We chose the K-means clustering as our solution. The idea of the K-means algorithm is to cluster the k clusters in space as the center, classify the objects near them, and then update the centroid values continuously through an iterative method until the optimal clustering is obtained. Using K-means clustering, we could intuitively see what group our data can be divided into based on the features to get the characteristics of these customers. We used the sklearn package in python both for clustering and classified methods.

6 Results

According to the problem statement, our ultimate goal is to improve the lifetime value and increase the company's profit. In order to achieve this, we must first respond to what the customer likes and increase these kinds of service to customers. The second is to reduce the churn rate. So the most important results of our project are: first, the result of finding the most valuable customers and finding out what they are interested in. The second is the result of how to reduce the churn rate because retaining old customers can make more profits than attracting new customers [1]. Last but not least is the ability to predict new customers' value.

In order to find the most valuable customers, we chose to analyze from three aspects. We first referenced the BG model in the *lifetime* package. The BG model is a probabilistic forecasting model designed based on the assumptions of the Pareto/NBD model. The BG/NBD model describes the repeat purchase behavior in the context of a non-contractual customer relationship. When the user is active, the number of transactions completed by a user within a time period t obeys a Poisson distribution with a mean value of λ_t [4].

By fitting this model, we drew two heatmaps of the customer's predicted number of purchases in the next 60 days and whether the customer is "survival." According to the heat map, we could see the relationship between the number of customers purchased and whether they "survive" and frequency and recency so that we could summarize the patterns of the most valuable customers from the perspective of frequency and recency.

Second, we found the most valuable customers from the lifetime value perspective. We focused on analyzing those customers whose purchase frequency is one or more. The reason is that attracting new customers costs companies more money and energy than retaining existing customers. Retaining existing customers is the most effective way to increase profits[1].We screened out these customers into a new data frame. We fitted a Gamma-Gamma model to the data[5].

This model looks at the correlation between average purchase amount and purchase frequency and calculates the possible future sales, predicted purchases, probability of being alive, and lifetime value for each customer. Through this method, we could get customers with the most significant lifetime value, their predicted purchase times, and purchase amount within two months.

The third aspect is customers with a high ratio of total revenue and recharge. The best customers are those who recharge more money and use this money in a short time. Therefore, we calculated the ratio of the total revenue they brought to the company to each customer's recharge amount, and the larger the ratio, the better. The k-means Clustering model was the option we chose here. The idea of the K-means Clustering algorithm is to cluster the K points in the space as the center, classify the objects close to them, and update the value of the cluster center (centroid) successively through an iterative method until the best cluster is obtained. The process of the K-means Clustering algorithm is as follows: First, select the center points of k categories. For any sample, find the distance to the center of each category and classify the sample into the category with the shortest distance center; after clustering, recalculate the center point position of each cluster; repeat steps 2 and 3 until the positions of k center points remain unchanged or a certain number of iterations is reached, then the iteration ends. Otherwise, continue to iterate.

According to the ratio, we were finally divided into four groups. We drew a violin plot for each variable and used the Pearson Correlation model, showing the correlation of every variable to each other, and we focused on the correlation between total revenue and other variables. The calculation formula of the Pearson correlation coefficient is as follows:

$$Pearson = \frac{Covariance\ of\ x\ and\ y}{standard\ deviation\ of\ x * standard\ deviation\ of\ y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-y)}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

When the correlation coefficient is 0, the X and Y vectors are not correlated. When the value of X increases (decreases) and Y decreases (increases), the two vectors of X and Y are negatively correlated, and the correlation coefficient is between -1.0 and 0.0. When the value of X increases (decreases), Y increases (decreases), and the two vectors of X and Y are positively correlated, and the correlation coefficient is between 0.0 and +1.0.

After finding the group with the highest ratio and viewing this group's relevant correlation and clustering information, we got what aspects and dimensions we should target.

For the second part of the results, reducing the churn rate can be inferred from observing the preferences of people with frequency=0, that is, people who only purchased once.

The last is to predict whether the new customer is a high-value or low-to-medium-value customer. We used logistic regression.

The output variable of this model always ranges between 0 and 1. The assumption of the logistic regression model is: $h_\theta = g(\theta^T X)$, among them: X represents the eigenvector, and g represents the logistic function, which is a commonly used logistic function as the Sigmoid function, the formula is $g(z) = \frac{1}{1+e^{-z}}$[6].

We divided customers into three groups according to ratio, customers in group 0 are the least valuable, and those in group 2 are the most valuable. We divided 80% of the data into the training set and 20% into the test set. The accuracy obtained by fitting the Logistics Regression model into our data is 80%.

I will now explain the specific results, graphs, and tables.

6.1.1

Our first step was to analyze the relationship between the number of purchases with the frequency and recency and the probability that the customer will not churn within 60 days. Figure 1 shows the heatmap of the expected number of future purchases for 60 days by frequency and recency. From the image, we can conclude that the customers in the yellow area in the lower right corner are the customers with the most purchases in the future. Therefore, customers who make repeated purchases often (high frequency) and have ample time between the first and the latest purchase (high recency) might become the best customers in the future. Furthermore, customers in the dark blue area in the upper right corner may never return. That is to say, customers who have made many repeated purchases (high frequency) but are within a small time interval between the first and most recent purchase (small recency) are likely to be lost. Last but not least, customers in the light blue and green areas are intrigued because they may purchase again, and we can expect them to make ten or more purchases within the next 60 days. These customers may need a little persuasion or nudges to purchase more.

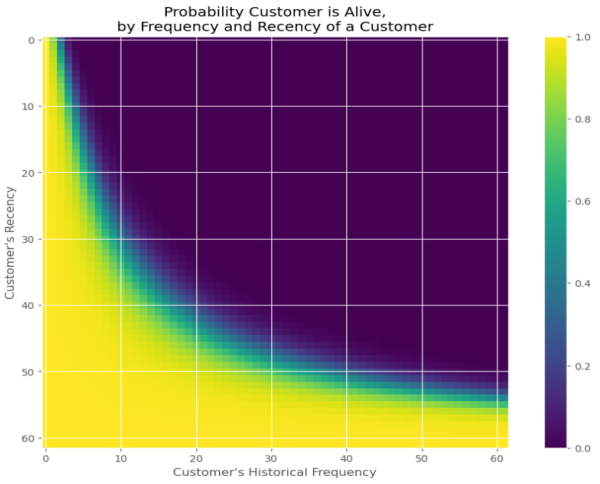

Figure 1 Heatmap of Expected Number of Future Purchases

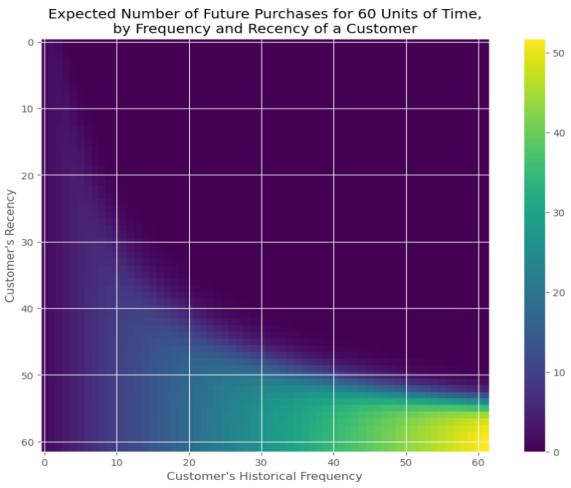

Figure 2 Heatmap of Probability of customer is "alive"

Figure 2 shows the probability that customers will not be lost in the next 60 days; that is, they will still be "alive." This heat table shows that the customers in the lower left and bottom yellow areas have almost one stock probability. For customers in this yellow area, regardless of whether customers frequently purchase, as long as their recency is greater than 55, that is, the time interval between the first purchase and the latest purchase exceeds 55 days, the possibility of their "survival" is almost 1.0. Moreover, customers with low recency, even if their frequency is low, that is, they do not make repeated purchases have a better chance of being alive. On the contrary, customers with many purchases but low recency are more likely to churn, which is in the large blue area in the upper right corner.

6.1.2

According to what I mentioned above, we applied the gamma-gamma model to calculate features such as lifetime values and churn. We decided to use 0.6 as the threshold of whether the customer will churn. If the probability is less than 0.6, the customer will churn. If the probability is between 0.6 and 0.75, the customer is marked as a "high-risk" user. If the probability is more significant than 0.75, the customer will not churn. Figure 3 shows the top 10 most valuable customers with alive rates, monetary value, and lifetime value (LTV). We stored all customers' information in a data frame and saved it as a CSV file. Besides, we counted the number of customers with churn, not churn, and high risk, as shown in Figure 4.

MSISDN	frequency	recency	T	monetary_value	predicted_purchases	p_alive	churn	predicted_Sales	
2202505230	25.0	58.0	61.0	224.560000	22.300006	0.994982	not churned	223.943164	7277.24
829815402	11.0	51.0	56.0	442.272727	11.343547	0.995755	not churned	438.851342	7285.36
2992759090	9.0	55.0	59.0	564.444444	9.100558	0.997051	not churned	558.884153	7450.14
8354830438	1.0	1.0	4.0	500.000000	11.174586	0.996990	not churned	460.161955	7514.42
7925821143	36.0	60.0	60.0	163.611111	31.958681	0.998434	not churned	163.356714	7578.91
85953282	25.0	61.0	61.0	240.480000	22.377340	0.998433	not churned	239.797611	7819.47
1365247585	8.0	49.0	54.0	612.000000	8.874368	0.996584	not churned	605.146918	7866.81
2145146746	1.0	11.0	15.0	1000.000000	5.887264	0.997577	not churned	913.320483	7880.84
2213792708	5.0	53.0	61.0	1040.000000	5.413744	0.996605	not churned	1020.449386	8103.86
1682453146	55.0	58.0	61.0	122.036364	46.156774	0.981287	not churned	121.947673	8122.70

Figure 3 Table of Predicted Values, i.e., Purchases, Churn, Lifetime Value, etc

```
total number of customers: 81664
not churned: 80737
churned: 496
high risk: 431
```

Figure 4 Number of customers with churn or not churn

6.1.3

Now we moved to analyze the existing customers with the k-means clustering model to divide them into different groups according to their characteristics. We combined their two-month expenses in various fields to get a new data frame. For the k-means Clustering method, we selected Total Revenue, Total Voice Revenue Total, Total Data Revenue, Total SMS Revenue, Active Days, Price Plan, Connection Type, Handset Type, Charged Data Usage, and Outgoing Minute Usage,' these 11 variables. By fitting the k-means model on our data, we got that k is equal to 5, as shown in Figure 5, which means that data can be divided into five groups according to different characteristics.

Figure 5 k Value Choosing

Using a violin plot to draw the feature distribution of the five groups, as shown in Figure 6, we analyze the characteristics of different groups from 11 aspects.

According to Figure 6, we can see that almost all of the ratios of customers in group 3 are 1, and the number of customers in group 3 is the most. So, we think our client should enlarge the scale of the third group.

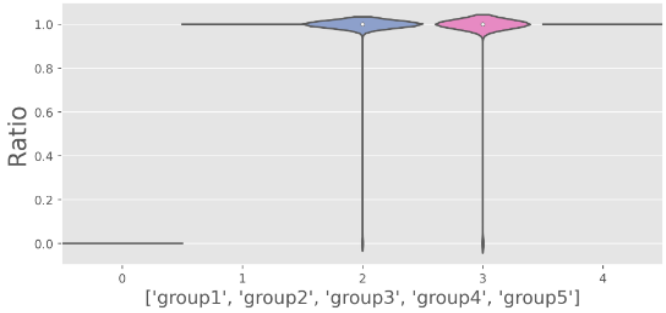

Figure 6 Violin Sub-Plot for Ratio

From Figure 7,8,9, we can get the characteristics of the third group of users. The third group of customers is relatively active (as seen from the active days). This group of customers uses data more than voice and SMS and is more inclined to use a 2G network. Moreover, the favorite price plans of this kind of customer are types 0,3,5, corresponding to BONDHU, DJUICE_ADJUSTED, and NISHCHINTO. The preferred mobile phone is 3, which is SMARTPHONE 4G.

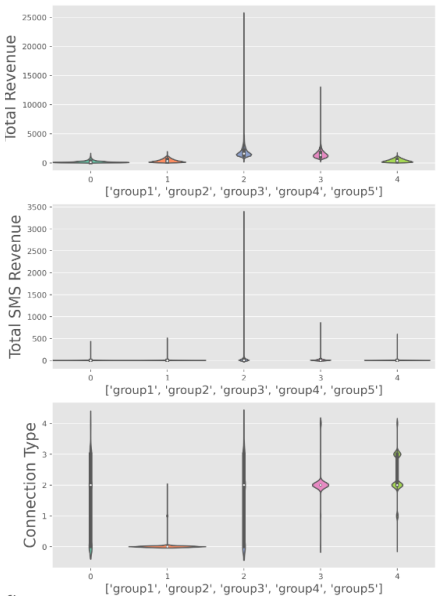

Figure 7 Three Violin Sub-Plots for Ratio

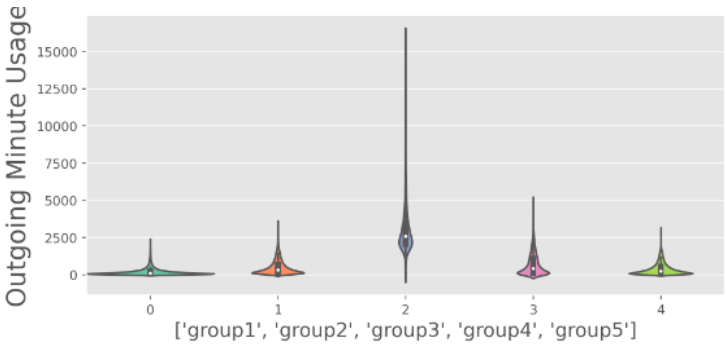

Figure 8 Volin Sub-Plot for Outgoing Minute

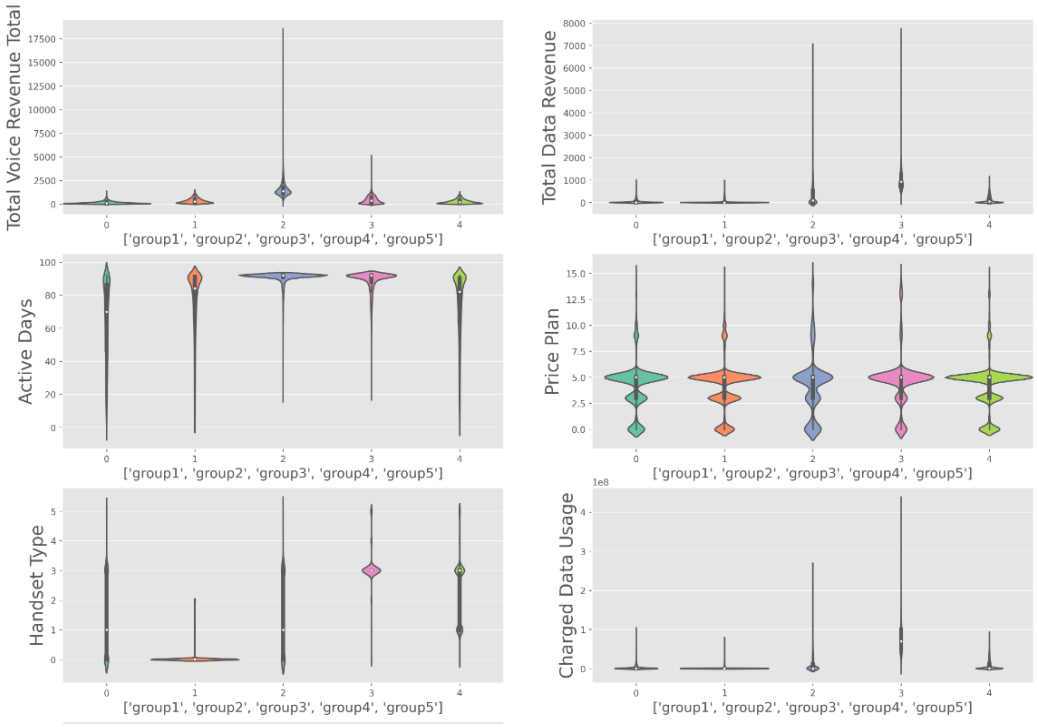

Figure 9 Six Volin Sub-Plots for High Freq Customers

Next, we obtained the data frame of the correlation between the variables of each group through Pearson Correlation. For the targeted third group of customers, total revenue is highly correlated with Voice revenue, data revenue, Charged Data Usage, and Outgoing Minute Usage.

	Total Revenue	Total Voice Revenue Total	Total Data Revenue	Total SMS Revenue	Active Days	Price Plan	Connection Type	Handset Type	Charged Data Usage	Outgoing Minute Usage
(3, Total Revenue)	1.000000	0.678659	0.756382	0.169362	0.260501	-0.036968	-0.045988	-0.032209	0.471714	0.520401
(3, Total Voice Revenue Total)	0.678659	1.000000	0.049093	0.229037	0.218148	-0.082145	-0.068493	-0.056217	0.050544	0.854775
(3, Total Data Revenue)	0.756382	0.049093	1.000000	-0.050233	0.162709	0.016709	0.000408	0.007017	0.612169	-0.040783
(3, Total SMS Revenue)	0.169362	0.229037	-0.050233	1.000000	0.060973	-0.068369	-0.032615	-0.019420	-0.020816	0.219578
(3, Active Days)	0.260501	0.218148	0.162709	0.060973	1.000000	-0.023310	-0.020737	0.023009	0.095160	0.187940
(3, Price Plan)	-0.036968	-0.082145	0.016709	-0.068369	-0.023310	1.000000	0.311140	0.267963	0.022040	-0.060944
(3, Connection Type)	-0.045988	-0.068493	0.000408	-0.032615	-0.020737	0.311140	1.000000	0.782246	0.057675	-0.040375
(3, Handset Type)	-0.032209	-0.056217	0.007017	-0.019420	0.023009	0.267963	0.782246	1.000000	0.060714	-0.028527
(3, Charged Data Usage)	0.471714	0.050544	0.612169	-0.020816	0.095160	0.022040	0.057675	0.060714	1.000000	0.074980
(3, Outgoing Minute Usage)	0.520401	0.854775	-0.040783	0.219578	0.187940	-0.060944	-0.040375	-0.028527	0.074980	1.000000

Figure 10 Correlation of Each Variable to others

6.2

To reduce churn, we paid attention to those customers whose frequency=0, that is, customers who have only purchased once. We figured out where the company can improve its services by looking for the differences in violin plots compared with the former ones. The customers with a frequency of 0 can be divided into four groups according to the k-means clustering model. Figure 11 shows the distribution of every variable of different groups.

According to the comparison with the previous customers with high purchase frequency, it was found that there is a significant difference in that more people in group 3 used 4G networks, and many people bought price plan 13, which corresponds to KHULNA. The handset type they used is Type 5, and this one is marked UNKNOWN. Therefore, our suggestion to the company is to improve the service quality of the 4G network, reduce the use of the price plan KHULNA, or upgrade this plan.

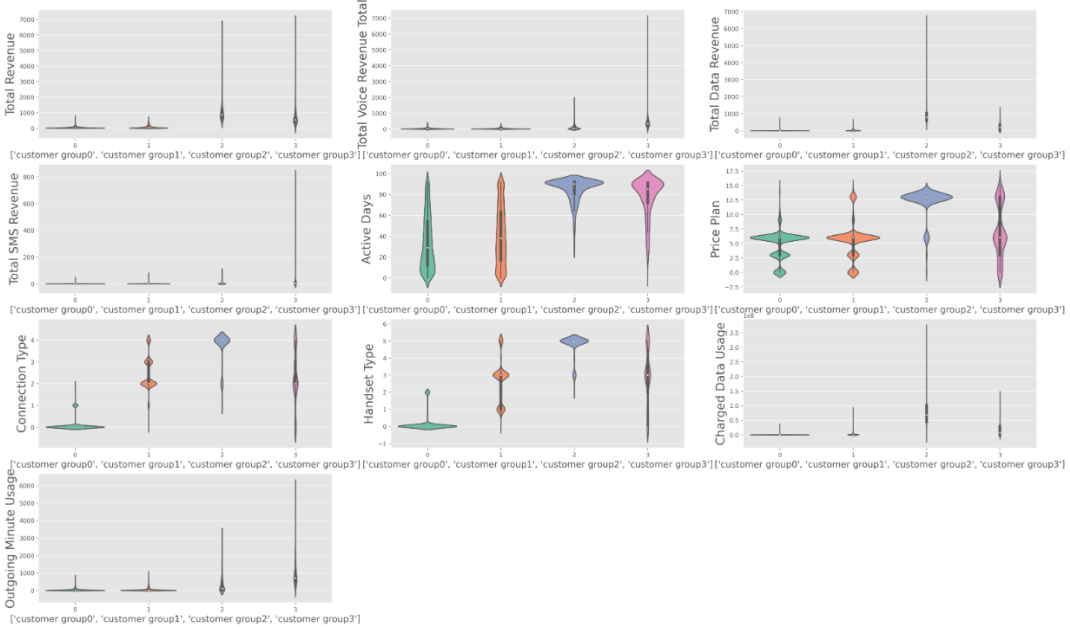

Figure 11 Volin Sub-Plots for Low Freq Customers

6.3

Last but not least, it is to predict which group of customers the new customers belong to. We divided them into three groups according to the ratio. Customers in group 2 are the most valuable, and those in 0 are relatively the least valuable. Figure 12 is a portion of our grouped data frame. We divided new users through logistics regression. The resulting model has an accuracy of 72.39%.

	Total Revenue	Total Voice Revenue Total	Total Data Revenue	Total SMS Revenue	Active Days	Price Plan	Connection Type	Handset Type	Charged Data Usage	Outgoing Minute Usage	Ratio
0	1.38	1.38	0.00	0.00	10	NISHCHINTO	4G	SMARTPHONE 4G	0.00	1.53	0.0
1	11.96	11.96	0.00	0.00	15	NISHCHINTO	BASIC	BASIC-VOICE ONLY	0.00	16.90	0.0
2	20.57	19.32	0.75	0.50	4	DJUICE_ADJUSTED	4G	SMARTPHONE 4G	769.14	9.53	0.0
3	58.59	54.59	0.00	0.00	43	NISHCHINTO	2G	BASIC-DATA CAPABLE	1.13	78.25	0.0
4	37.63	37.63	0.00	0.00	8	NISHCHINTO	2G	BASIC-DATA CAPABLE	1.12	22.10	0.0
...	...	...	...	...	...	...	...	...	...	...	...
81604	741.47	335.78	405.67	0.00	84	DJUICE_ADJUSTED	4G	SMARTPHONE 4G	29937059.73	486.08	2.0
81605	297.75	297.75	0.00	0.00	76	NISHCHINTO	BASIC	BASIC-VOICE ONLY	1.78	507.72	2.0
81606	288.92	129.27	158.66	1.00	69	PREPAID SMILE PSTN	3G	SMARTPHONE	8530344.97	177.48	2.0
81607	572.11	275.92	162.72	133.45	80	DJUICE_ADJUSTED	4G	SMARTPHONE 4G	8130163.68	338.43	2.0
81608	119.92	119.42	0.00	0.50	52	DJUICE_ADJUSTED	BASIC	BASIC-VOICE ONLY	0.00	81.84	2.0

Figure 12 Portion of Different Groups of Customers

7 Limitations

Our model assumption has been based on the data that our sponsor provides. However, the shifting preferences of clients are always active. The new technology development or competition from other companies may influence consumer choices, and we cannot predict it. For this project, The data we analyze was nearly three months. Our model was referential and accurate for the analysis and prediction of recent data. But we couldn't ensure that our model and analysis kept the accuracy for future data or the data from a long time ago. Therefore, we recommend that our sponsor retrain the model with new data at regular intervals.

In addition, our team's selection of variables to fit the model was limited and biased. There are up to 50 different variables in our sponsor's data, many of which are unknown meanings and can not be transformed into clear data. Our team considered these variables and decided to ignore them. However, we can not guarantee that these variables are related to total revenue or customer lifetime value. It is a very challenging thing to determine the relationship between each variable. Also, our sponsor has provided a lot of data, which added to the difficulty of our building model and training it. Therefore, we can not currently solve the problem of fitting all variables into the model. However, our team believes that as we learn more, we would be able to find new algorithms and models to analyze these variables in the future.

8 Interpretation of Results

We divided the entire dataset between lower frequency purchasing and higher frequency purchasing in ratio. By applying K-mean clustering, we got that k was equal to 5, which generated five groups of people. According to the violent graph, the third group had the highest ratio, meaning this group is more valuable, and the company should enlarge its services and take action. Generally speaking, when the group's ratio is high, they also have higher frequency and recency in customer lifetime values. By digging

more deeply, the correlation of each factor will be assessed to determine which factor plays more effectively. The purpose of determining the most influential group is also to uncover related characteristics, such as plan types, model types, how much data values, and voice values they spend. Therefore, we found out that total data revenue played an important role with total revenue as our Y variables. Also, we noticed that people in group 3 are likely not interested in using SMS services. Retaining new consumers is much more complicated and earning profits compared to returning buyers. Regardless, our suggestion to the company is to continue offering various data plans to reach customers' demands.

On the other hand, we also care about customers with low purchase frequency. According to the second violent graph, it demonstrated that those people used 4G networks more often and bought different price plans. Although only a small number of people were considered churn, we think the company still wanted to minimize the number of customers lost in the churn. So, we recommend that the company can improve the service quality of the 4G network, update prince plans, or even have promotions on posting advertisements to attract more customers. In terms of maximizing customers' lifetime value, we think the customer lifetime value model is a new area of investigation. It best shows the probability that customers retain and profit in a long-term relationship. We have clarified that there are some limitations on the accuracy of the models for future prediction. Modern technology is changing rapidly. We want to recommend retraining the model with different data or even different approaches to fit into the new dataset monthly.

9 Individual Contributions

9.1 Lilly Pan

Technical Contribution: Lilly Pan handled data visualization and partial math modeling, which produced crucial patterns and results. She used the ggplot2 package in Python to develop the visualization result and the BG model and Gamma-Gamma mode to evaluate customer lifetime value, which will help our team classify customers.

Nontechnical Contribution: She presented the Midterm presentation, took notes when had a meeting with our sponsor, and contributed to all the Tech memos and presentation slides. She also did a lot of research online and helped the team resolve technical concerns.

9.2 Cathy Tang

Technical Contribution: Cathy Tang used the approach of removing records with missing values for data cleaning and used the *Pandas* package to combine two primary datasets for math modeling in further study. She also separates and outputs the data based on the customer's lifetime value level. Finally, determine the relationship between total revenue and a set of dimensions by using a correlation test.

Nontechnical Contribution: She presented Elevator Pitch, communicated with our sponsor, formatted Gantt charts in Tech memo3, and contributed to all the Tech memo and presentation slides.

9.3 Helen Zhu

Technical Contribution: Helen Zhu took the most responsibility for math modeling. She applied the k-mean clustering and identified the number of k clusters by SSE to partition the data into different groups, which determined characteristics and preferences for our client's products. She

also tried the logistics model, which would help our sponsor to predict their new customer's CLV(customer lifetime value) level.

Nontechnical Contribution: She presented the Midterm presentation, presented the Tools and Techniques Presentation, emailed our sponsor and communicated with them, and contributed to all the Tech memos and presentation slides.

10 Conclusions and Future work

Throughout the entire project, we have arrived at a few main conclusions. Due to the large dataset, we spent a lot of time on data cleaning to remove all unnecessary columns and rows.
We chose the Gamma-Gamma model to evaluate the correlation between average purchase amount and purchase frequency. We pursued mainly to target higher frequency and recency people. Therefore, we separated data by ratio and split it into different levels. Moreover, we also used k-mean clusters as one approach, which allowed us to do classification based on each characteristic and feature. Our results show that group 3 was the most valuable group when the frequency is larger than 0 in our violent graph. This can be a good reference for future customers if they want to make a suitable marketing investment.

In the future, we want to study further what other factors affect recency. Recency typically is the number of days between repeated purchases. Therefore, we may finalize a way to know why churned people leave the market. Furthermore, since the accuracy of our model is only 80% after splitting the data into two sets, we plan to keep working on it in order to improve the accuracy for business uses. To complement the analysis, We would use active days and regions to evaluate the churn rate and, based on other classification factors, to see which factors are related to churn if we have a chance to tackle this program again.

“We have neither given nor received unauthorized assistance on this assignment.”

11 Works Cited

[1] Flordal, P., & Friberg, J. (2013). Modeling customer lifetime value in the telecom industry. Lund University.

[2] Razak, N. I., & Wahid, M. H. (2021). Telecommunication customers churn prediction using machine learning. 2021 IEEE 15th Malaysia International Conference on Communication (MICC). https://doi.org/10.1109/micc53484.2021.9642137

[3] Seippel, H. S. (2018, March 1). Customer purchase prediction through machine learning. Retrieved October 15, 2022, from https://essay.utwente.nl/74808/

[4] *"counting your customers" The easy way: An alternative to the pareto ...* (n.d.). Retrieved December 8, 2022, from http://brucehardie.com/papers/018/fader_et_al_mksc_05.pdf

[5] Hardie, B. G. S. (n.d.). *[PDF] the gamma-gamma model of monetary value: Semantic scholar*. [PDF] The Gamma-Gamma Model of Monetary Value | Semantic Scholar. Retrieved December 7, 2022, from https://www.semanticscholar.org/paper/The-Gamma-Gamma-Model-of-Monetary-Value-Hardie/8e3408bab082f7b81ad5bd5b35e7dc48f1859f82

[6]Sperandei, S. (2014). Understanding logistic regression analysis. Biochemia Medica, 12–18. https://doi.org/10.11613/bm.2014.003