

Using Clustering to Analyze Repeat Callers

Michael Barnhart
Hood College
401 Rosemont Ave
Frederick, Maryland 21701
Meb26@hood.edu

ABSTRACT

This paper uses a clustering data mining approach to analyze repeat callers for a prepaid credit card queue over the length of a month. However, due to limitations of the data set, a more simplified approach will be done using only date info and interpreting info from that. It will be used to determine time and day of most calls. Using Cluster ID to find times that certain lines of business must find peak times for better employee scheduling and possible service pricing for peak times. Using these techniques, it will be found that the time of highest repeat call volume is during 11:00AM CST-1:00PM CST while the day with highest repeat call volume are Fridays. The day of the months plays no statistical role in the amount of repeat call volume, but would require more data to verify this claim.

Categories and Subject Descriptors

- Information systems→Data Mining;
Clustering;

Keywords

Data Mining, Call Center Study, Repeat Caller, Clustering.

1. INTRODUCTION

In the ever-increasing competitive market of our service economy, companies are looking for ways to decrease their costs, and increase the productiveness of their Customer Care Organizations (CCO)s [6]. There are several ways in order to do this, but this paper will tackle the aspect of repeat callers and trying to identify possible causes if those repeat callers. Methods that will be used for this analysis is

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Conference '04, Month 1–2, 2004, City, State, Country.

using clustering algorithms to detect patterns of the repeat callers. These patterns would open insight as to the cause of those repeat callers to allow for a solution to be more easily formulated to eliminate those repeat callers.

2. Repeat Callers in CCO

2.1 Identifying Repeat Callers

To determine how to handle repeat callers, repeat callers need to be properly identified and labeled using a set number of criteria [3]. The data that is being used is a repeat caller report for a phone queue for supporting prepaid credit cards. The parameters for this data are that to be a repeat caller, someone must call in at least three times within a three-day period. The third and subsequent call ins will be added to the list. The entries will come with the time of that repeat call, the Line of Business (the client that supports that prepaid credit card), phone number that they are calling in on, and an ID to ID that call. The list is compiled at the end of the day to be seen the next day. Most of this can be seen in the in Table 1.1. The phone number has been removed as that is Person Identifiable Information as is not allowed to be given out as per the Payment Card Industry.

Table1.1: An excerpt of the data being used once the first round of data cleaning has been done to remove personal info.

Record_Number	RecordWritten	Cluster_ID	Repeat_UID
3782448	10/14/2018 1:26	2	73793717- 0eb3-44ad- 97a0- 3047d8e09152

2.2 Why Repeat Caller are Important

Cutting costs are a huge focus of Customer Care Organizations and major cost issue is Repeat Callers [2]. Repeat Callers cost at least twice as more as those who had First Call Resolution as they must call in multiple times to

get an issue resolved instead of the first time [9]. Repeat Callers end up becoming former customers as they become frustrated from having to call in multiple times for the same issue, taking their money with them. Furthermore, Repeat Callers lead to bad PR as customer will voice their frustrations to outside sources such as social media which can put affect how potential customers see the company leading to future possible missed losses.

2.3 Strategies to Handle Repeat Callers

Ways to handle repeat callers involve identifying repeat callers so that Second Call Resolution can be accomplished [4]. If the call can be addressed in that second call, callers are less likely to call in again, costing the Call Center less. One proper identification is set up the Customer Care Organization will need to set up procedures to better handle Second Call Resolution. However, making procedures takes time and it is in the best interest of the Customer Care Organization to tackle that efficiently [6]. Creating a way to determine reasons for Repeat Callers will allow the Customer Care Organization more efficiently handle them.

3. Prepaid Credit Cards

To better understand the call types, it should be explained what a Prepaid Credit Card is. A Prepaid Credit Card works similarly to a Debit Card, but instead of pulling money from a checking account like a Debit Card, Prepaid Credit Cards have the money preloaded on the card [10]. The main users of Prepaid Credit Cards are those who do not have a bank account, but still need to have their paychecks cashed in a convenient way. Prepaid Credit Cards are not like gift cards or banking accounts with their own fees and terms and conditions that can be found here: <https://www.consumerfinance.gov/ask-cfpb/category-prepaid-cards/>

4. Datamining Approaches

4.1 Clustering

Due to the limited dimension of the data set clustering is the most useful option for determining patterns in classical datamining techniques [5]. Clustering would allow for determining areas of high call activity for each line of business over a date range. Furthermore, clustering will also allow for greater visibility into the what this data look like in this circumstance of having limited data sets.

4.2 KMeans

KMeans will be the clustering algorithm used for clustering the data of the repeat callers in a efficient way. KMeans is able to handle larger datasets more efficiently than a hierarchal approach which is needed for this project as the dataset will include entries that number in the tens of thousands [8]. Using KMean, it should be able to reliably create accurate clusters of the repeat caller data. Other clustering methods were looked at for the run of this project that are listed under the scikit-learn python library [1]. However, it was decided that the KMeans algorithm is able to handle the job in an efficient matter and produce the desired result that is we are looking for, so KMeans is the algorithm of choice for this project.

4.3 Other Datamining Methods

Due to the dimensionally small dataset, other datamining methods would not work well at identifying any new information or perform well with data that is available. However, using more traditional statistical models and graphs will be used as the data is in a format that lends itself well. Checking several aspects wit in regards to time can found as to time of day and day of the month which show the most volume which can be used to identify peak times and days for scheduling purpose of the phone agents, to which line of business brings in the higher call volume which could be used in price negotiations with lines of business that bring in higher traffic than others.

5. The Dataset

The unfortunate truth of the simple dataset sets back the possibilities of what can be done. It also makes it more difficult to determine reasons for the repeat callers as this data is not listed and must be inferred from other sources such as ticketing system or pay schedules of the card owners' employers. This maybe remedied if the repeat callers could be integrated with data from the ticketing system, or some other system that can capture intents of calls such Speech Analysis, it would conceivable to be able to determine much more information from this data versus if its not there. More info that could be found would be the price of each of these calls that cost he business money. With this extra information it could be found what call types have the highest financial burden so that the business can identify them and possible try to reduce or even eliminate the number of repeat callers due to that issue. As of the writing of this paper the data set is being worked on

to hopefully add in these additional figures so that this project would be a more fruitful endeavor.

6. Datamining Results

6.1 Days in the Month

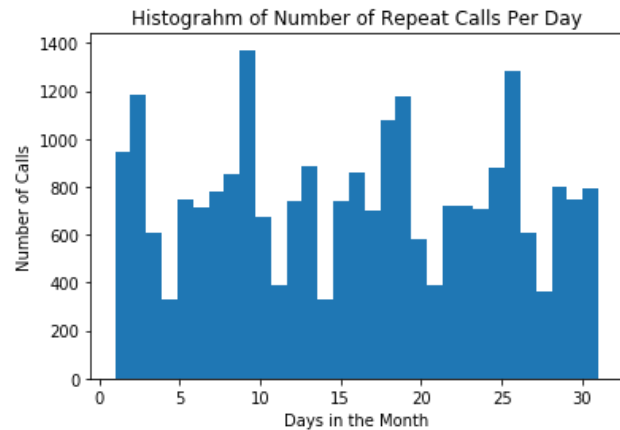
The first thing to be looked at is how repeat call volume looks on a per day basis of the month. The first thing to be done is to preprocess the RecordWritten portion of the original data. The issue is that the date in that data field is in a format that matplotlib python library cannot read. So to fix this issue the date data under RecordWritten was changed to only reflect the day of the month for that call. So for example if the date was “10/14/2018 1:34 PM” it is changed to only show “14”, a much friendlier value for the library to work with. This was done to every line of data to properly show what day of the month that his data was used for.

Table 1.2: An excerpt of the new data used for the “Days of the Month” part analysis only showing the day of the month, under RecordWritten, and the Cluster_ID referring to which LOB the repeat call went to.

RecordWritten	Cluster_ID
14	2
14	3
14	4

After the data has been trimmed to show only the necessary data for analysis, see Table 1.2, the data was running through python library matplotlib to create a histogram of the to visualize the days of the month that experience the highest amount of repeat call volume versus other days, see Figure 1.3. It was found that the day of highest repeat call volume was on the 9th day of the month, 11/9 to be exact, at 1368 repeat calls coming on that day alone. However, while this does show that the 8th has the highest value overall, the reason for that high number is most likely not because of it being the 9th day of the month, but rather that the day in question is a Friday. Looking at the histogram it can be seen that the volume of repeat calls seems to follow a trend. Looking at the Figure 1.3 the days with the highest repeat call volume are on Fridays (10/19, 10/26, 11/2, 11/9). A look into how the day of the week plays into the effect of repeat call volume will be investigated next to see if the relationship stands.

Figure 1.3: A histogram of the repeat call volume on a per day basis for days in a month



6.2 Days of the Week

Similar to Days of the Month, the date data under RecordWritten had to be preprocessed in order for it work with the matplotlib library. To accomplish this the date was used to find out what day of the week the date corresponds to. So every day that was a Sunday is labeled as a Sunday, every Monday will be a Monday etc. This, however, was not enough as the library does not work well with strings. The days had to be converted to integers to play more nicely. So for this each day of the week was assigned an integer ranging from 1 to 7 for Sunday through Saturday respectively. This allowed for the data to be easily read for graphing purposes which can be seen in Table 1.4.

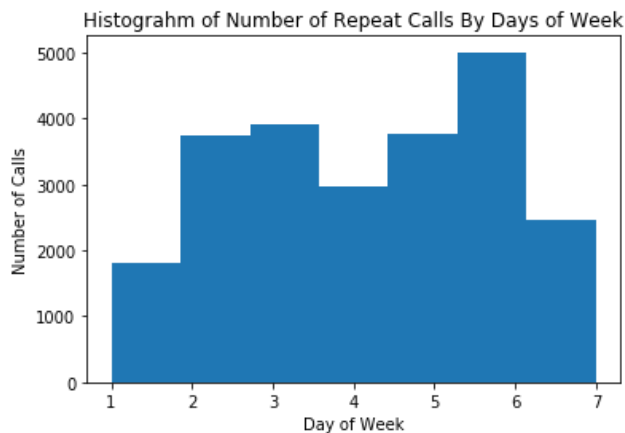
Table 1.4: An excerpt of the data used for “Days of the Week” part of the analysis only showing the day of the week, here 1 meaning “Sunday” and the Cluster_ID showing the LOB the repeat call went to.

RecordWritten	Cluster_ID
1	2
1	3
1	4

After the data trimming the data was ran again through matplotlib’s histogram function to create a histogram of the data, shown in Figure 1.5. From the histogram it can be

seen that the data follows a similar trend that was previously seen in the “Days of the Month”. The trend being that repeat call volume seems to increase going from Sunday to maxing out on Friday and then dropping off dramatically on Saturday. The possible reason for this trend is that people tend to get paid closer to the end of the week, usually Thursday or Friday. The average user would most likely call on these days to see what their balance is on their card to see if their check went through and will most likely call back to see it’s in yet. The sudden drop off on weekends is an interesting case to look at. This could be explained by people simply not wanting to call in on their day off, or not knowing that the line is open on weekends, which is the case with a lot of call centers. However, these claims are outside the scope of this research and will require additional time and resources to find out the reason for this. Overall, it can be said that repeat call volume hits its highs on Fridays while dropping off for the weekend. Weekdays experience on average more per day repeat call volume then the weekends.

Figure 1.5: A histogram of repeat call volume by days of the week



6.3 Time of Day

For Time of the Day the Record Written portion had to be edited once more. The “Date” portion of the data had to be removed to leave only the time of day the repeat call came in. 24-hour time had to be used be used as the library is not capable to understand the string portion of “AM” or “PM”. Then once the data was in the 24-format (00:00) the colon was replaced with a decimal to allow it to work with the library properly. This also allows for it to graphed using the 24-hour timeframe vs a 2400-time scale used in military time purposes though both would be able to accomplish the

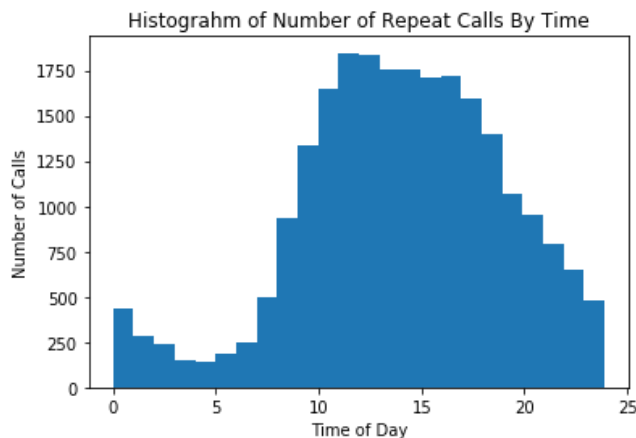
task. The data was trimmed down to only include the time and the Cluster_ID, as seen in Table 1.6. All time data is using Central Standard Time as a base.

Table 1.6: An excerpt of the data used for “Time of Day” part of the analysis only showing the day of the week, here 1.26 means “01:26 AM” and the Cluster_ID showing the LOB the repeat call went to.

RecordWritten	Cluster_ID
1.26	2
1.31	3
1.33	3

Using the newly trimmed data a histogram was made of time of day repeat calls using the matplotlib library. The data was split into 24 sections to represent the 24 hours of the day. The histogram shows the volumes for each hours of the day. From this graph the peak time for repeat calls is in the 11:00 AM to 1:00 PM timeslot with average of over 1750 repeat calls per hour as seen in Figure 1.7. The trend then starts to decay there after and continues to do so until 5:00 AM, where the number of repeat callers starts to climb back up to the peak at 12:00 PM. From this it can be deduced that repeat call volume tends to happen during the daylight hours instead of nighttime. The reason for this is most likely because most people are up at these hours of the day in the United States, the call center does not handle non-domestic calls for this call type. It can also be seen that the rate in which the repeat call volume increases at a much faster rate then it decreases. The reason for this phenomenon would need to be researched further to elucidate the reason, but a possible reason could do when they wake up versus when they go to sleep as people tend to wake up at the same time but go to sleep at different times. Another reason is that people are more willing to try to call again in the earlier hours of the day versus the later hours as they have more time to call again to accomplish what that set out to do, but this claim would need to be researched more to give a proper answer.

Figure 1.7: A histogram of the number of repeat callers based on the time of day that they call in



7. References and Citations

In conclusion, after isolating the data to examine repeat call volume by a day of the month, day of the week, time of day it was found that repeat callers most often call in during the 11:00AM CST-1:00PM CST timeframe and during Fridays. Furthermore, it was found that day of the month seems to play no real bearing on repeat call volume, but that would require a larger data set to see how day of the month data plays a role in repeat call volume.

Using the information found it can be said that people usually call in during normal waking hours looking at time of day. Peaking out at lunch time on the east coast and slowing decaying. The call center would best utilize its workforce by scheduling to have most of its agent during the 12:00 EST time clock. With these being repeat calls it would also be best advised to have the most supervisors during this time as well as repeat calls tend to have a higher chance of supervisor calls vs first close calls as people are less willing to take answers that don't accomplish their need. With repeat calls happening mostly during Fridays the call center should have their agents have Friday included in their schedule. The normal Monday to Friday workweek does cover this, however agents' schedules do not always follow this, especially if said agents work the weekends.

For future work in this research it would be advised to change the data set to include the reason for the repeat call. Including the reason for the call would allow the call center to better identify the major causes in repeat call volume.

With this understanding the call center can either create solutions to better handle the issues of these repeat calls, and/or improve existing technologies to handle the issue before the call reaches the agent. This will not only improve the call experience for the caller, but also save money in the call center for not having to staff as many agents during peak times to handle the repeat call volume. First Call resolution solutions are the best thing for the call center to use to have the greatest benefit of customer satisfaction and cost savings.

8. ACKNOWLEDGMENTS

Thanks to my boss for allowing me to use this data and in supporting me in the project as well as my friend, the bio major, for berating me into writing better scientific papers because in her words "are not even that hard".

9. REFERENCES

- [1] Anon. 2.3. Clustering. Retrieved October 22, 2018 from <http://scikit-learn.org/stable/modules/clustering.html>
- [2] Serina Aswani. 2017. 5 Ways to Cut Costs in the Contact Center. (August 2017). Retrieved October 22, 2018 from <https://www.clarabridge.com/blog/5-ways-cut-costs-contact-center/>
- [3] Robert Bushney, Benjamin Anthony Knott, and Sarah Korth. 2009. System and method of determining call treatment of repeat calls. (December 2009).
- [4] Neil Crane. 2010. Strategies for Handling Repeat Calls. (October 2010). Retrieved October 22, 2018 from <https://multichannelmerchant.com/marketing/strategies-for-handling-repeat-calls>
- [5] Saurav Kaushik. 2016. An Introduction to Clustering & different methods of clustering. (November 2016). Retrieved October 22, 2018 from <https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/>
- [6] Carl Krumins. 2018. How To Reduce Call Centre Operating Costs Overnight. (January 2018). Retrieved October 22, 2018 from <https://thehub.smsglobal.com/reduce-call-centre-costs>
- [7] Pedregosa et al. 2011. Scikit-learn: Machine Learning in Python. (October 2011). Retrieved October 22,

2018 from

<http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html>

- [8] M. Steinbach, G. Karypis, and V. Kumar. 2000. A comparison of document clustering techniques. *KDD workshop on text mining* 400, 1 (2000), 525–526.
- [9] Andrew Studee. 2012. Expert's Angle: Rethinking Repeat Callers. (December 2012). Retrieved October 22, 2018 from <https://www.icmi.com/Resources/Learning-and-Development/2012/12/Rethinking-Repeat-Callers>

- [10] Anon. 2017. What is the difference between a prepaid card, a credit card, and a debit card? (March 2017). Retrieved October 22, 2018 from <https://www.consumerfinance.gov/ask-cfpb/what-is-the-difference-between-a-prepaid-card-a-credit-card-and-a-debit-card-en-433/>