

Information as Power:

Implementing Data Analytics at CARD Bank



Grameen Foundation, a global nonprofit organization, helps the world's poorest people – especially women – lift themselves out of poverty by providing appropriate financial services, life-changing information and unique business opportunities. Founded in 1997, Grameen Foundation has offices in Washington, DC; Seattle, WA; Colombia; Ghana; Hong Kong; Indonesia; Kenya; the Philippines; and Uganda. Microfinance pioneer Dr. Muhammad Yunus, founder of Grameen Bank and winner of the 2006 Nobel Peace Prize, is a founding member of its Board of Directors, and now serves as director emeritus. For more information, please visit grameenfoundation.org.

This case study is a product of Grameen Foundation. We would like to acknowledge Kate Druschel Griffin as the writer of this case study; Jacobo Menajovsky and Linda Gonzalez for technical support; and Debbie Dean and Julie Peachey for additional support. We would like to further acknowledge CARD Bank for their partnership and collaboration.

©2012 Grameen Foundation® USA, All Rights Reserved

Except for use in a review, the reproduction or utilization of this work or part of it in any form or by electronics, or other means not known or hereafter invented, including xerography, photocopying, recording, and in any information storage, transmission or retrieval system, including CD ROM, online or via the Internet, is forbidden without the written permission of Grameen Foundation.

Introduction

Grameen Foundation's Microsavings Initiative, a three-year project funded by the Bill & Melinda Gates Foundation, is working to enhance access to safe, flexible, convenient deposit accounts for poor households by building poverty-focused savings programs at microfinance institutions.

Grameen Foundation works with CARD Bank, Inc., a microfinance-oriented rural bank in the Philippines, which has offered voluntary savings products in the past but has not achieved the desired outreach or portfolio volume. Grameen Foundation is assisting with market research and product development, technology improvements, human resources, treasury management and social performance management to help create a more robust savings approach for CARD Bank. One fundamental piece of this is utilizing the data CARD Bank generates on its clients – information about who their clients are and how they save – to help them understand financial behaviors, reach more savers and better serve clients with relevant and tailored savings products.¹

This case study illustrates the use of data analytics – including the use of the Progress out of Poverty® Index (PPI®) – to strengthen CARD Bank's savings strategy. It outlines the business questions we asked, the client insights we gained, and how this is being used to change product design and delivery. We also supply a step by step analytical process so readers can see how this was achieved.

Data Analytics

Grameen Foundation believes it is critical to not only observe how poor households manage their money, but also understand how households with different personal attributes – such as income levels, educational levels, ages or family make-up – might manage their money in different ways. This allows an institution to observe patterns, make predictions and perceive how client behavior might change with adjustments to product terms or delivery channels. Using information such as clients' total savings amounts or the types of accounts they prefer, institutions can make better informed decisions about product strategies to target specific segments of their clientele. Examples of these strategies include marketing a particular product attribute over another to clients below a certain poverty line or creating financial literacy programs tailored for a targeted group.

¹ CARD Bank maintains proper confidentiality and privacy of its clients in compliance with regulatory requirements and Grameen Foundation worked with CARD to maintain that privacy in this project.

I. Business Questions

It is best to begin this type of work with a set of business questions that should be analyzed during the process. (Another option is to outline a set of hypotheses to test with the data, as Grameen Foundation did in our data analytics work with [PT Ruma](#).)

In our financial services work at Grameen Foundation, we use a generic set of business questions to analyze customer and transaction level financial data (detailed in the tables on the next page). These business questions are at the customer and the product level, as outlined below. Grameen Foundation's senior data analyst Jacobo Menajovsky summarizes our approach best saying, "This is not magic....People mainly just open and close accounts, deposit, withdraw and borrow money. We just need to track frequency, amounts and channels used, and most importantly, check if any of these is affected by poverty or some other socio-demographic indicators."

We used these questions to help us understand a few things:

- *What was CARD's poverty outreach in its savings portfolio?*
- *Which other CARD products are they accessing?*
- *Are we missing opportunities for cross-selling?*

Overall, we wanted to have a better understanding of client characteristics and needs to help improve CARD's marketing strategies, and in turn better serve clients with new or updated savings products and services.

Customer Analytic Roadmap



Business Question	Analytical Approach	Details
What is the average customer probability of living below a certain poverty line?	PPI	Measure probability of customer being below poverty lines, and track progress over time.
Do some customers cluster in a particular way? Do they share a particular set of characteristics?	Customer Segmentation	The PPI offers a good set of indicators to segment customers, such as "family size", "kids in school", "educational levels", "type of employment", etc. These types of segmentations are simple and can help develop new products or services.
How much do my poorest customers contribute to the business?	Volume Contribution (revenue)	The overall contribution at different poverty levels.
	Transactional Contribution (could be a cost or a revenue center depending on transactional fees)	
How much time do our customers need before stabilizing their savings balances? Does this look different at different poverty levels?	Balance Stabilization Trends	Checking customer average balances among customers with the same tenure will provide a better understanding of how much time it takes to stabilize their balances.
Do customers show different "transactional" profiles? Does this correlate with their poverty level?	Inflows vs. Outflows Profiling (frequency & amounts)	Customer indicators to capture financial behavior based on the frequency and the amount of transactions.
How do customers use the available delivery channels? Does this differ by poverty level?	Delivery Channel Profiling (frequency & amounts)	Indicators on how customers use the available delivery channels to access products and conduct transactions.
Do customers show a specific customer lifecycle?	Customer Lifecycle	Main characteristics that define the average customer lifecycle.
How do customers earn their money?	Earn It Profiling	Type of declared businesses and income generating activities.
Do customers show different "saver" profiles?	Protect It Profiling	Customer tendencies on how they use their savings account (to save or transact).
Do customers show different growth levels in their savings accounts?	Grow It Profiling	Indicators to capture savings growth levels.
How we can we rank/measure customer profitability?	Customer Profitability	Customers by the revenue they generate and how that relates to poverty movement.

Product Analytic Roadmap

Savings

Credit

Payments

Insurance

Financial Ed

Business Question	Analytical Approach	Details
Do (poor) customers have a different set of products?	Product Mix Mapping	Research into the different set of products that customers have and check if we find significant differences depending on their poverty levels and financial and socio-demographic indicators.
Do some customers have more products than others?	Cross-Sell Product Penetration	Main drivers for cross selling products and services.
Is there any particular order in which customers open their products?	Product Opening Patterns	Patterns of opening products and how those differ by poverty levels and financial indicators.
For savings: How are savings products being used? Does this differ by poverty levels?	To Keep vs. To Transact Product Profiling	Performance of savings products and how are they being used (to save or to transact).
For credit: How are credit products being used? Does this differ by poverty levels?	Length and Depth Product Profiling	Performance of credit product in terms of length and amount disbursed.
Do customers have distinct saving and borrowing patterns?	Saving and Borrowing Cycles	Indicators to capture the level of savings against borrowing behaviors.
Do customers have distinct credit patterns?	Regular vs. Sporadic Borrower Profiling	Customers that regularly are taking loans versus those with a more sporadic behavior.
Which should be the next product to offer to customers?	Next Product to Sell	Map out factors such as poverty level, demographic and financial information and stage in the customer life cycle to understand which should be the next product to offer to customers.

II. Data Sources

In conducting this business analysis, we accessed a range of data CARD collects about its clients and their financial behaviors.

Progress out of Poverty Index® (PPI®)

The Progress out of Poverty Index is a tool developed to measure the likelihood of clients living below a poverty level. Its ten indicators together form a score that predicts the poverty likelihood of the household. The PPI complements financial measures and allows an institution to use a combination of financial and social indicators to track client poverty levels, enabling it to assess its social performance and the effectiveness of its products.

In the Philippines, the PPI collects information² on the following:

- Number of school age children and whether they attend school
- Education level of the female adult in the house
- Existence of salaried employment in the household
- Construction of the house's outer walls and roof
- Type of toilet facility
- Ownership of refrigerators, television sets, and washing machines

CARD collects PPI data on all clients at each loan application, and when they open a savings account. We are able to use both the aggregate PPI score as an indicator, as well as the individual indicators mentioned above, to segment clients based on these characteristics.

CARD supplies Grameen Foundation with a group of clients' average balances on a monthly basis for most of their products. This means we have access to information on the following³:

- Latest PPI score
- Demographic information from individual responses to PPI questions
- Current number of products the client accesses
- Savings balances
- Outstanding loan balances

This meant we could see what was happening at the end of each month for each client in the sample. We did not have “transaction level” data for clients, however, thus limiting our ability to answer some of the business questions. For example, two clients with the exact same savings balances at the end of one month would have been viewed the same, when in actuality, one may have been transacting frequently

² More information on the PPI in general, and on the Philippines specific index, can be found at <http://www.progressoutofpoverty.org>

³ CARD does not supply any identifying information such as client's names, address, or government IDs

throughout the month while the other may have been inactive (not actively using the account). The data we had access to did not allow us to distinguish between those two clients.

We looked at data primarily from two of CARD Bank's most popular savings products, as well as from the product Grameen Foundation helped CARD to introduce, the Matapat account. The first is the Pledge Account (PA), which began as the compulsory account that CARD Bank customers are required to open when they begin their membership (take out their first loan). To obtain a subsequent loan, a customer must have 15% of the loan amount on deposit in the Pledge Account. In fact, many members are joining CARD just to save (they are not required to take a loan). Members earn 2% interest per annum on their balances in the account; deposits are made at the center meeting – the weekly meeting between CARD customers and field staff – making it a very convenient way to save. Withdrawals are allowed at nearby CARD Bank micro-banking offices (MBOs).

The second is the Kayang Kaya Account (KKA), a voluntary account with a PHP 500 (US\$ 11.83) minimum deposit and a 2% interest rate. Account openings, deposits and withdrawals can only be done at the branch office. The account was designed to target non-CARD members.

The third is the Matapat account, which is also a voluntary account. It can be opened at the center meeting in the presence of a bank officer, and deposits and withdrawals can happen at either the branch office or via an ATM machine. In addition, customers can sign up to have a savings agent collect deposits at a place of their choosing, including “on demand” via SMS. The Matapat account is being piloted by CARD Bank currently and is offered at three of CARD Bank's branches.

Table 1: Savings Accounts and Balance, as of April 30, 2012

	Number of Accounts	Balance (PHP)	Balance (USD)
Kayang-Kaya Account	19,011	182,702,600	4,248,898
Pledge Account	605,933	1,554,447,547	36,149,943
Matapat (3 branches only)	12,996	16,093,394	374,265

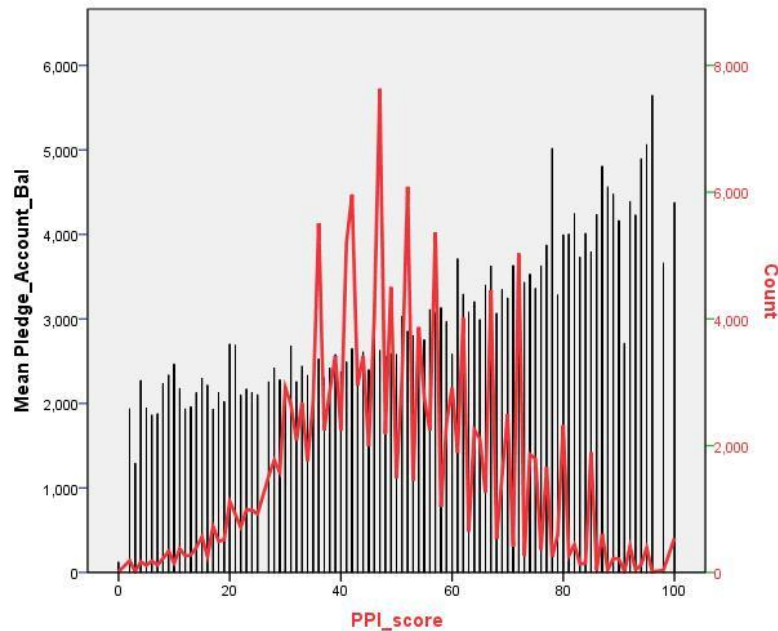
III. Findings

In Figure 1, we outline the major findings that resulted when we applied our business questions. The results are not mapped out in a neat, linear way to the business questions – as we analyzed the data, we were able to go into interesting and unplanned places, following what the data was telling us. Most data analytics processes are similarly messy.

Initially, we reviewed the natural customer segments within the CARD client population to help us analyze product mixes and the ways different clients contributed to the business differently. We asked, were clients with similar financial behaviors clustered together in the dataset? Do certain customers access one product more than others? How do customers' contributions to the business compare across segments? Customer segmentation is a significant investment of time and energy as it is often as much art as science. However, it can yield important insights into how different types of clients are responding to products in very different ways.

We also explored whether poverty levels could tell us anything about customer behaviors. The first thing we saw was a difference in the overall savings balances of poorer and less poor clients – the least poor clients were saving roughly 150% more than the poorest clients. Figure 1 shows average account balances (in the Pledge Account) increasing as PPI scores increase. The red line shows the distribution of poverty levels among the entire sample. However, there was no significant correlation between PPI scores and average account balances⁴ – which means that being poor did not necessarily mean you would be saving a lower amount. This has potential implications for the business case of serving very poor savers.

Figure 1: Poverty Level and Savings Balances



⁴ Pearson correlation= .146

We also examined whether people at different poverty levels responded differently to the type of delivery channel being offered. The Matapat product we worked with CARD to pilot allows a customer to deposit money both at center meetings and branches and to withdraw at a branch or an ATM. We found no significant difference in the percentage of money transacted (although transaction size did vary) or the delivery channel based on poverty levels. Figure 2 below shows this by poverty quintile – with the top 20% being the least poor and the bottom 20% the poorest households. This is a useful insight as the operations team discusses which delivery channels to emphasize over others.

Figure 2: Poverty Level and Delivery Channel

Poverty segments based on PPI score			
	% savings kept	% withdrawals thru Branches	% withdrawals thru ATM's
Top 20%	4%	11%	85%
2	7%	7%	86%
3	5%	13%	83%
4	6%	9%	86%
Bottom 20%	6%	6%	88%

For the segmentation process, we used PPI indicators to help us segment clients by certain characteristics such as family size, type of employment, etc. For example, client responses about the number of people below 14 years of age were re-coded to segment clients into different categories: big families, small families, and families with no children. The responses to questions about children's school attendance were categorized as children not in school, children in school, and families without children of school age. This enabled us to transform the PPI responses into new variables that allowed us to observe and segment CARD clients more effectively for the remaining business questions.

We looked at product mixes to see whether we could learn anything by analyzing which customers were accessing which products and used this information to build customer segments. For example, we looked at the characteristics of clients who held both the Pledge Account (PA) and the Kayang Kaya Account (KKA) to see what distinguished a client who had only the PA versus a client who held both accounts. To understand those differences we first built a balanced sample of clients⁵ with a PA and clients with both the PA and the KKA.

We found that clients that shared certain characteristics were more likely to hold both accounts, namely:

- Salaried employment
- Higher level of education
- Membership for more than 2.5 years
- Savings balance of more than PHP 3000 (US\$ 70)

⁵ This is a methodological step before running the statistical model required; checking so many variables inherently will require statistical software and classification algorithms.

Cross-Selling the Kayang Kaya Account

The analysis of the customer segments is also being used to develop a new marketing campaign to help CARD Bank cross-sell its Kayang Kaya Account (KKA). We can apply the shared characteristics of the Pledge Account and KKA customers to the entire dataset to test whether the customer segmentation applies - do we find, in fact, that households with these characteristics have a higher probability of opening a KKA? Once verified, we can then build a marketing campaign for the KKA that targets customers with these characteristics. We anticipate a higher likelihood of being able to sell to these customers; tracking over time we can see if that hypothesis is correct and verify or dispute the segmentation model.

Another method of checking whether our customer segmentation of cross-sold KKA and PA holders is correct is to “plot and count” customers as a way to validate the model. We did this by analyzing savings balances and PPI scores. In Figure 3, almost 80% of account holders had more than PHP 3000 in total savings, with the distribution of these cross-sold customers being concentrated above a PPI score of 50¹ (less poor). This helps prove that our customer segmentation outlined above is correct: KKA holders are more likely to be salaried and more educated (both PPI indicators) and have a higher savings balance.

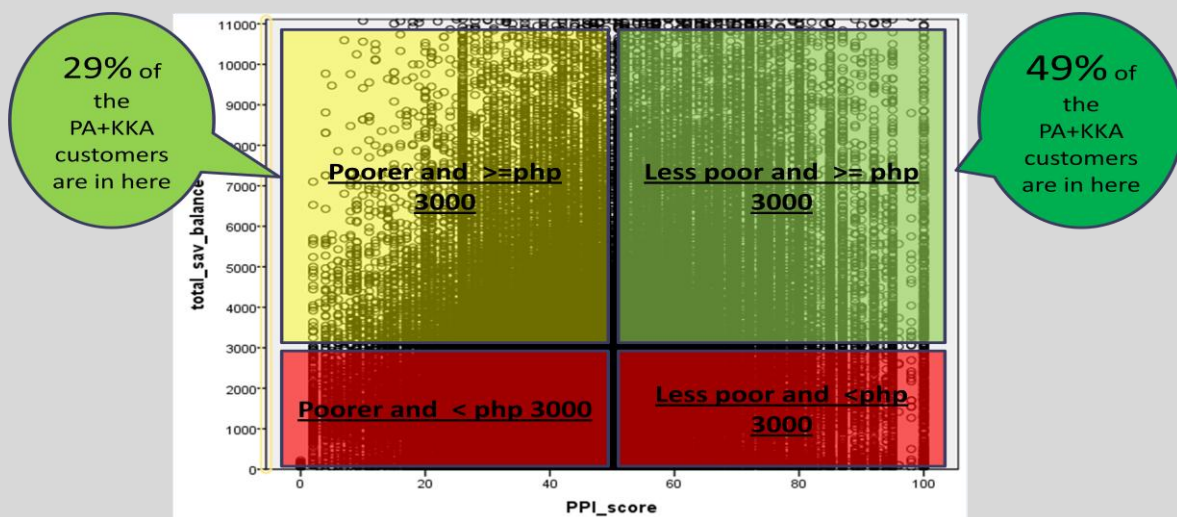
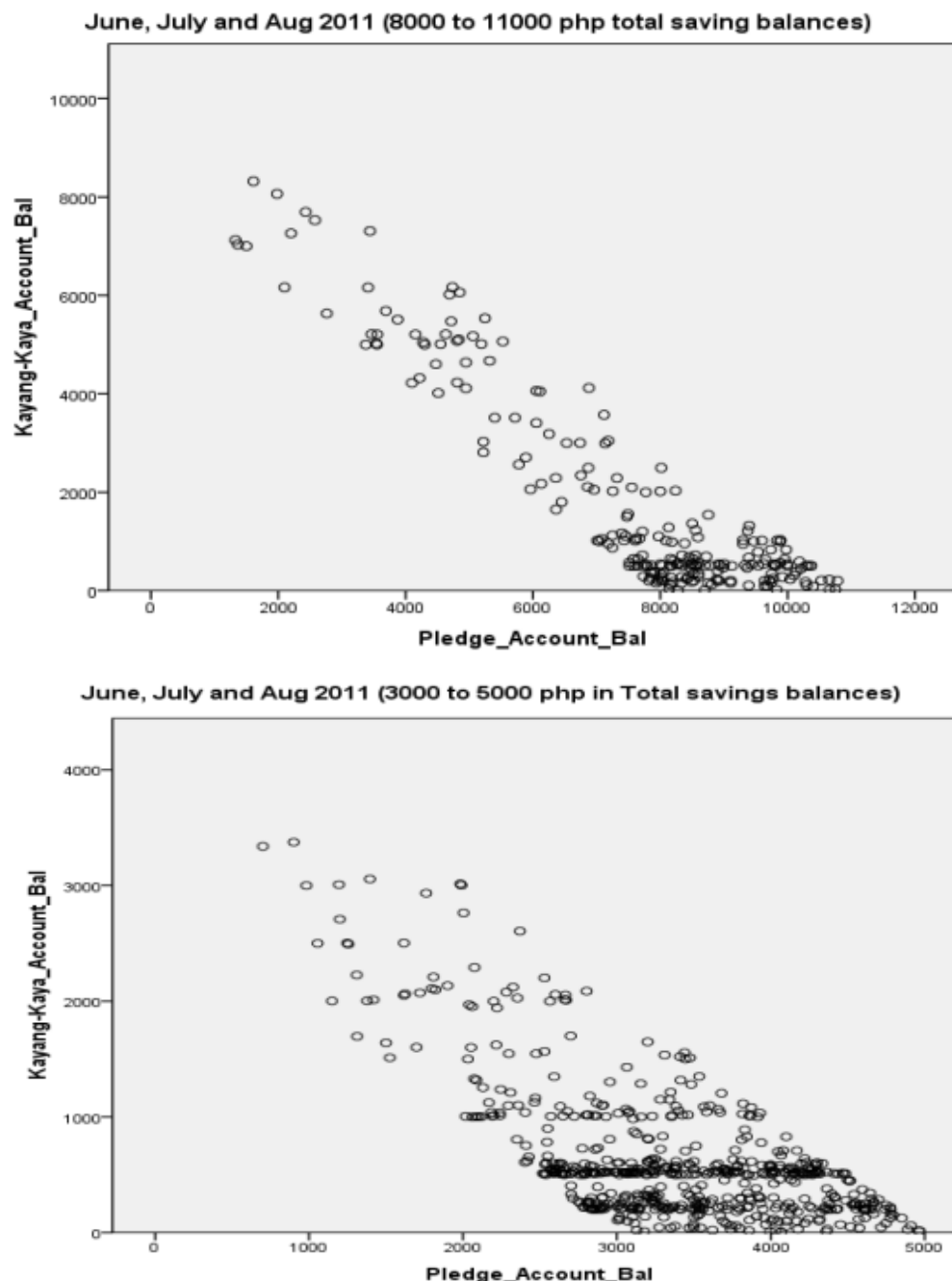


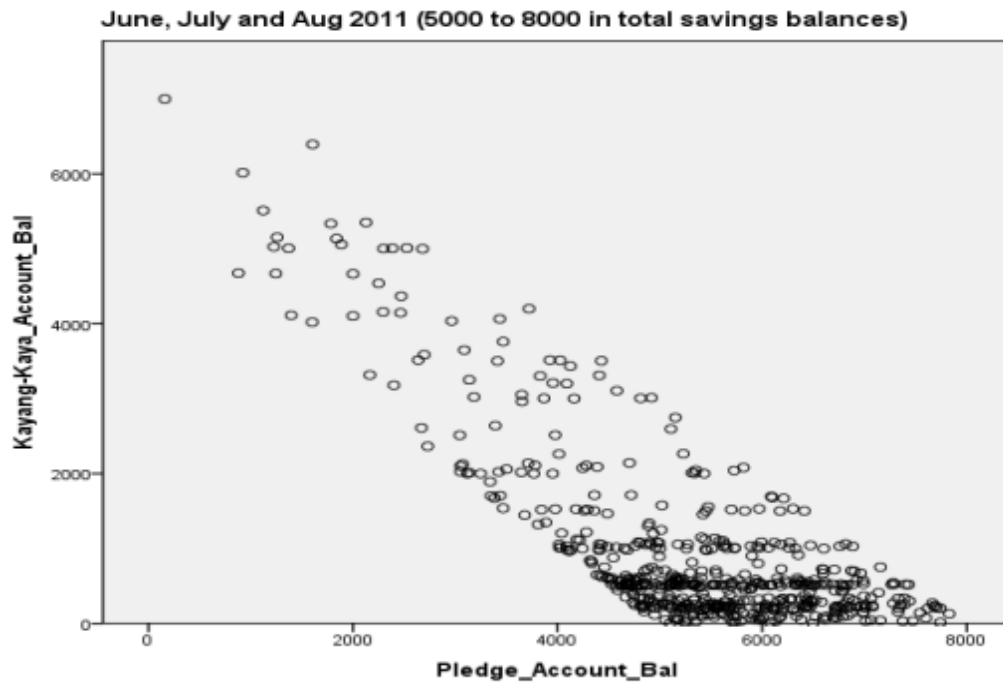
Figure 1: KKA client segmentation

We also found an interesting relationship between the Pledge Account and the Kayang Kaya Account. Clients with higher PA savings balances tended to have lower KKA balances (and vice versa). We segmented clients with a total savings balance of PHP 3000-5000 (US\$ 70 – US\$ 117), PHP 5000-8000 (US\$ 117 – US\$ 187), and PHP 8000-11000 (US\$ 187 – US\$ 257). As savings balances increased, the negative correlation between both accounts also increased. This means that clients who had higher PA savings balances tended to have lower KKA balances.

The following graphs are based on some of our findings in this analysis and clearly show that higher KKA balances correlated to lower PA balances, and vice versa – regardless of the total savings balance of the customer.

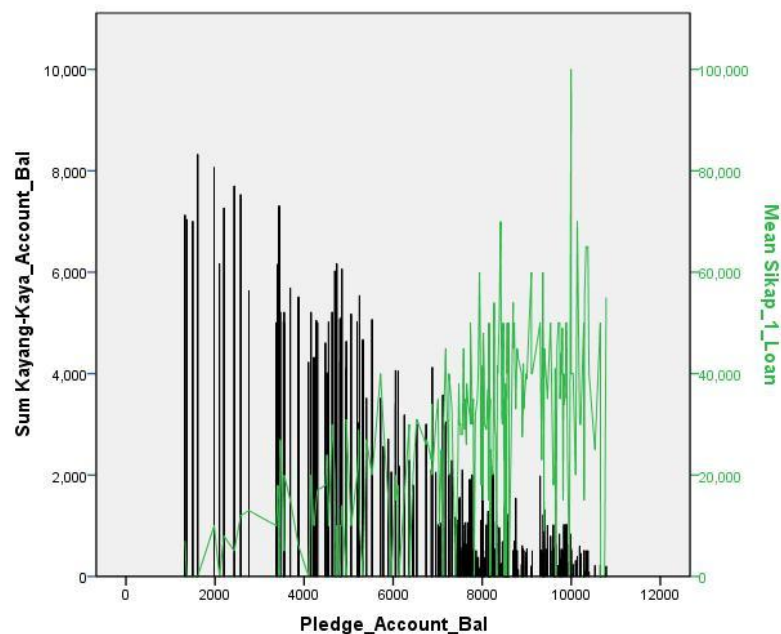
Figure 3: Comparing PA and KKA Balances





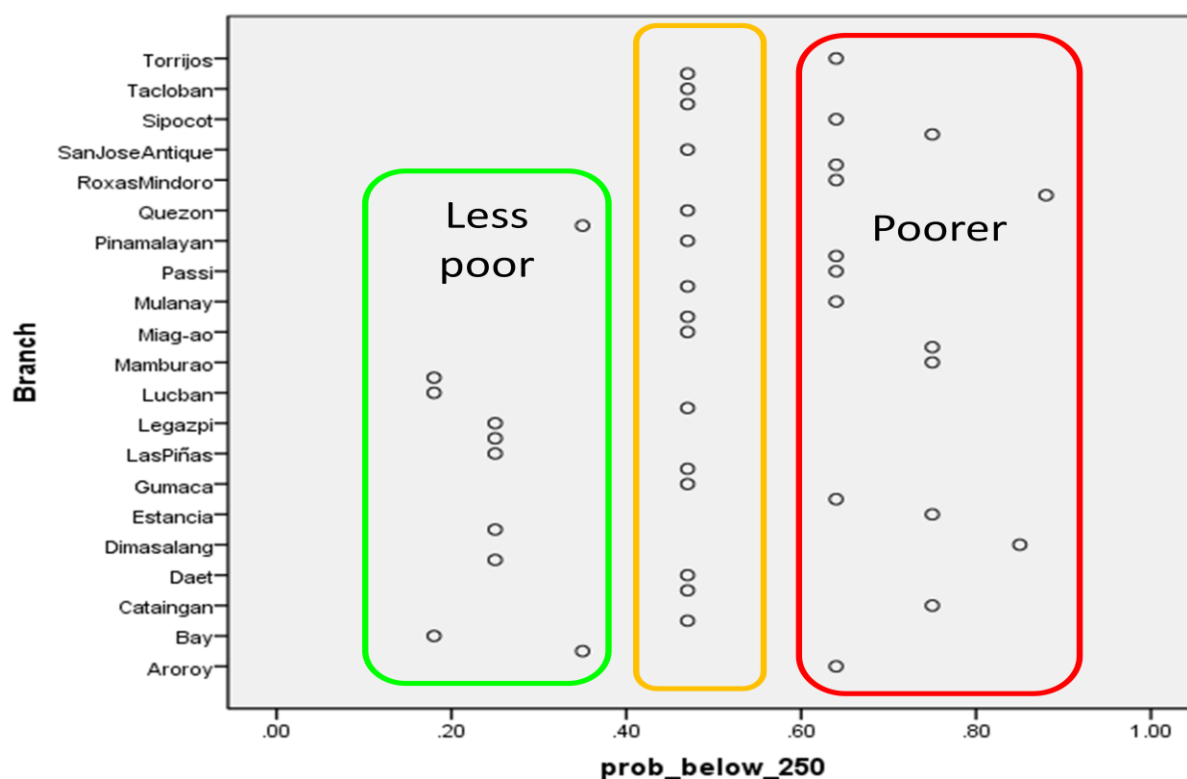
This led us to further ask, why are clients not saving more in the KKA? Is the PA preventing this? The PA requires a customer to deposit 15% of their initial loan amount as collateral against the loan, while the KKA is a voluntary savings account with no link to CARD's loan products. We then analyzed how savings behavior is associated with loan balances, using the hypothesis that a client could be forgoing depositing money in the KKA since she would need that money for collateral in the PA. As the graph in Figure 5 depicts, customers with high pledge account balances have low KKA balances (the black lines), while also having high outstanding loan balances. Customers with low pledge account balances have high KKA balances and low outstanding loan balances.

Figure 5: PA and KKA Balances vs. Outstanding Loan Balances



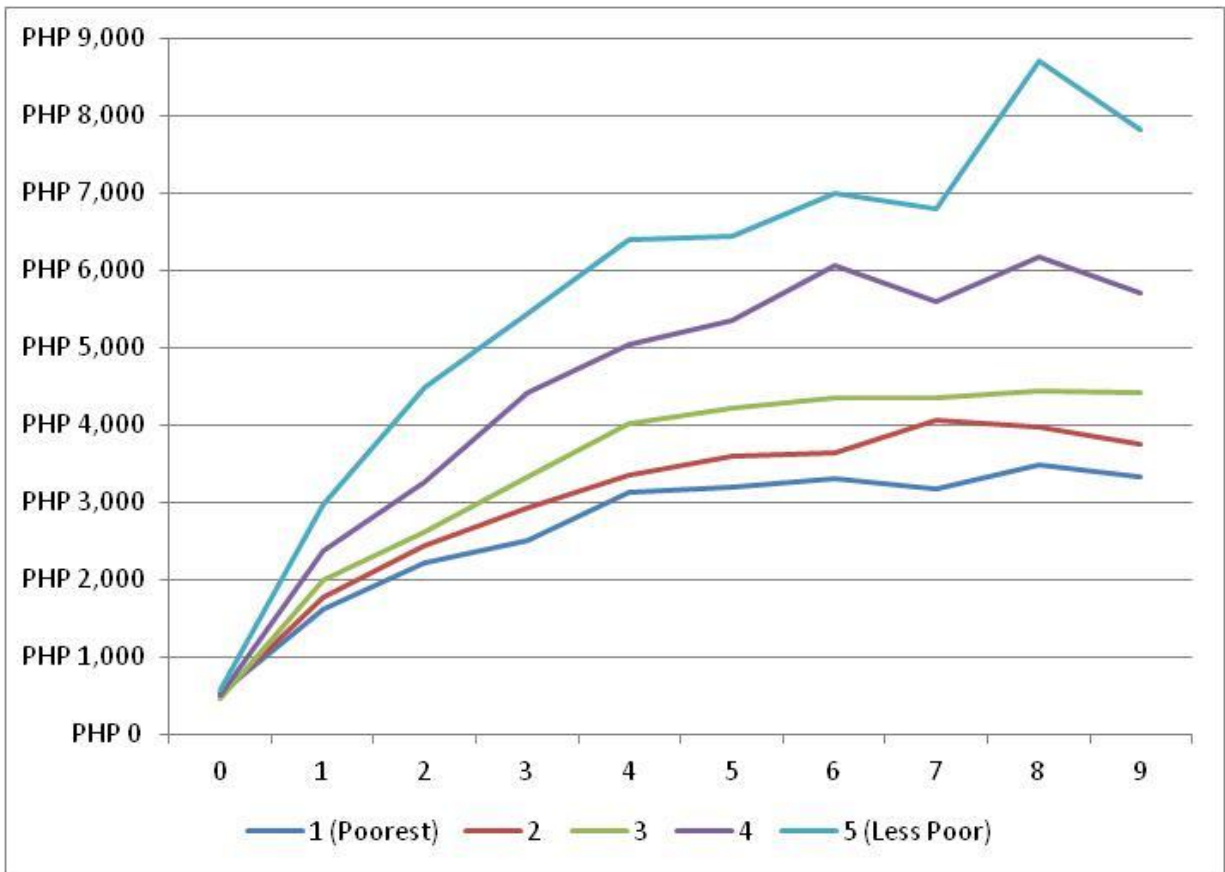
We are also using the PPI to examine poverty outreach by branch to help determine where we should pilot products if we want to target customers with a greater likelihood of being poor. In Figure 6 for example, we segmented CARD branches by poverty level. We can choose to pilot a low-cost minimum feature savings product to the poorer branches with specific range of poverty levels to make sure we reach the poorest clients.

Figure 6: Segmenting Branches by Poverty Outreach



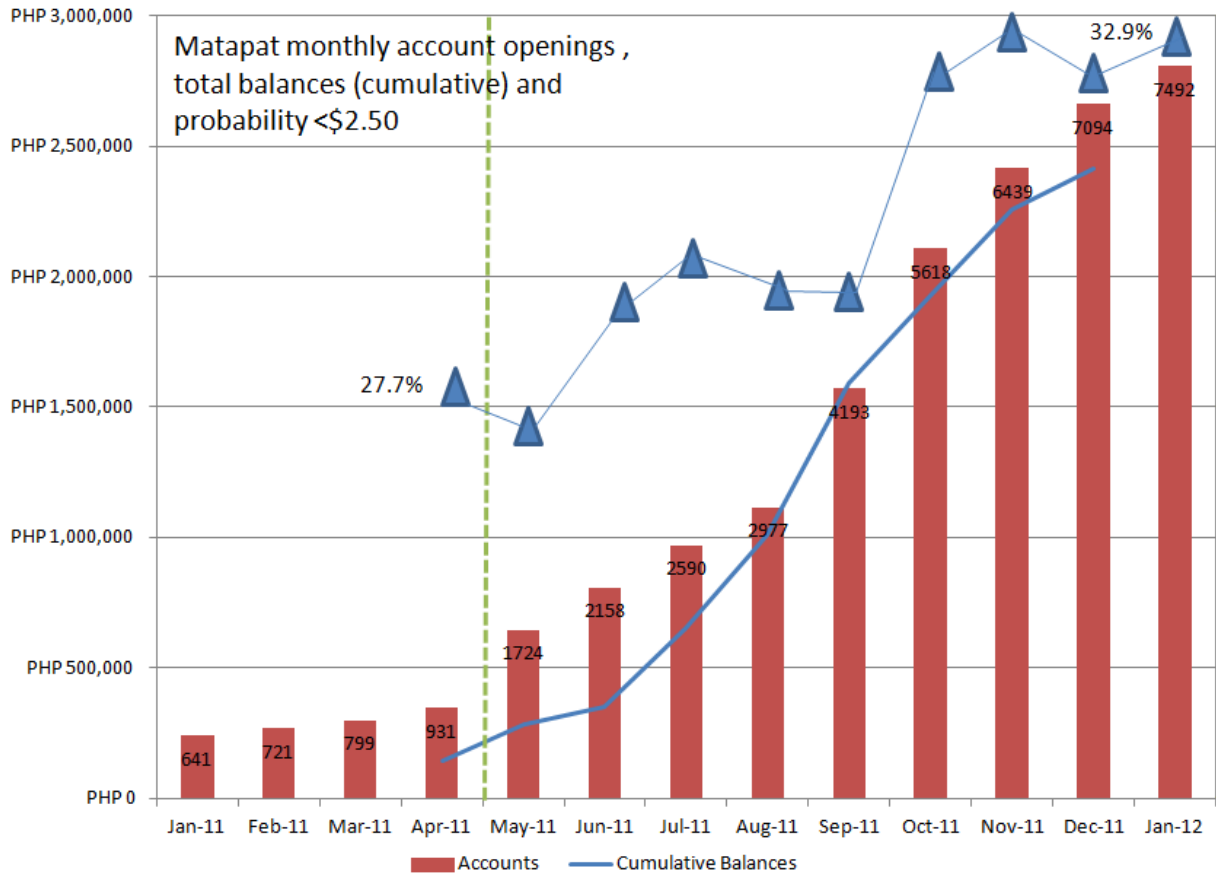
We also looked at customer tenure against savings balances to understand the possible behaviors of older customers with respect to increasing savings balances. We found, as outlined in Figure 7, that savings balances level off after an initial period of growth. CARD clients reach the average balance by year four of saving with CARD. We could use this information further for qualitative research on why this is happening and whether we should focus any marketing efforts on increasing the savings balances of longer-term customers.

Figure 7: Customer Tenure and Savings Balances



As we began piloting a new savings product, the Matapat, we were also able to use data to understand the effects of different product features. Initially, the minimum opening balance was set at PHP 1,000 (US\$23) – a figure that CARD management felt comfortable with from a business management perspective. In the first four months, there was low uptake of the account – fewer than 1000 accounts had been opened in that time. In April 2011, CARD dropped the minimum opening balance to PHP 100 (US\$ 2.30) and saw account openings increase significantly. We also saw poverty outreach deepen – from roughly 28% of account holders being under \$2.50/day to almost 33%. As we might have guessed, opening balances declined – but this is being offset by the increase in total number of accounts and the cumulative balances in this accounts.

Figure 8: Impact of Changes in Opening Balance



We are only just beginning to tap the potential of CARD's available data on clients. We look forward to continuing our work together to investigate transaction patterns and other information that will give us a broader picture of financial behaviors and product performance, as well as suggestions for improving on CARD's excellent customer service.

IV. Putting these Findings to Work

Since presenting these findings to CARD leadership and savings team, we have discussed the following actions for their current products, marketing efforts, new market research, and the steps they will take to institutionalize this type of data analytics going forward.

Product Portfolio Changes

The data analytics work revealed there was little cross-selling of products and possibly cannibalization of those products. This is mainly because product features are not very distinctive. The CARD Bank-Grameen Foundation team is currently refining the value proposition of each product in CARD's savings portfolio to help encourage cross-selling. The field officers now have a target for voluntary savings products in addition to the compulsory pledge account. Eventually, we will introduce targets specific to each product, and perhaps a target based on the cross-sell ratio.

As CARD roll-outs its Maagap (kiddie) savings account, it may be able to use the information from the PPI variables on family size for targeted marketing tactics to specific households.

Data Analytics in CARD's Future

Grameen Foundation delivered both this analysis and an in-depth training on data analytics and the use of the statistical software package SPSS to CARD's marketing and research teams. CARD aims to take on this kind of work in the future. For example, CARD's new marketing head is requesting customer profiles, and the team will use its new analytical skills to refine the segmentation. CARD's Research team now uses SPSS to tabulate results of its research studies, analyze monthly PPI reports, undertake random sampling when conducting market research, and to filter and clean data. These are all significant steps for the organization to integrate data analytics into its way of doing business.

Furthermore, CARD is currently upgrading its Management Information System to accommodate a core banking system that will provide integrated, transaction level detail reports. This will significantly improve its ability to conduct this type of data analytics going forward.

The CARD marketing team will ultimately be able to drive the future use of data analytics, as the organization makes better use of customer segmentation and product usage information to target products. For example, cross-selling of products is an area where CARD can improve, and where data can help to narrow field staff's focus on whom to target which products to. Product distinction will be another key component of CARD's marketing work. Data analytics will validate the success – and failures - of that work as CARD creates distinct marketing messages for different segments and develops value propositions for separate products.

V. Conclusions

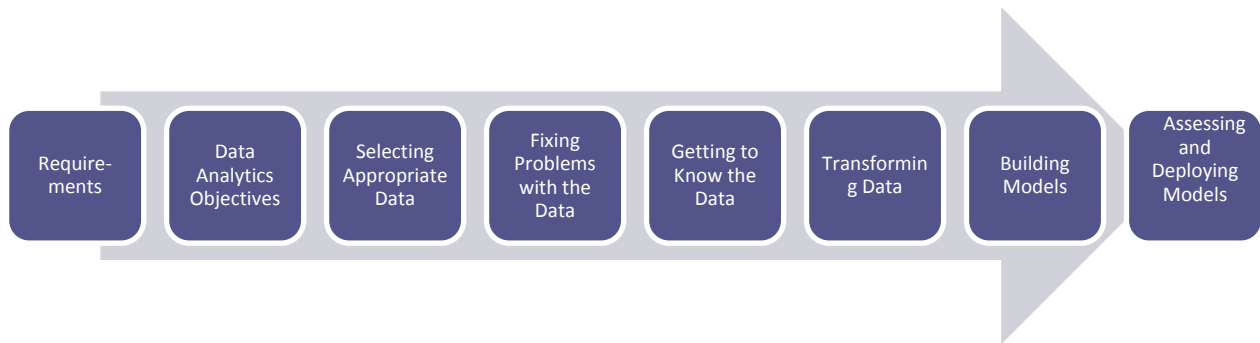
The field of microfinance is realizing the powerful practice of data analytics. This case study helped to showcase how Grameen Foundation analyzed data at CARD Bank to reveal new insights about their clients, savings behaviors, and poverty outreach for the institution. The annex provides more information about how an institution might approach a similar data analytics process.

Overall, Grameen Foundation advocates for social businesses to be very mindful of the power of data in helping to understand clients and lead to better design of products that are suited to their customer's cash flow management needs. We suggest the following:

1. In addition to the traditional demographic and transactional data most financial institutions have on customers, social businesses should also collect poverty data – through a practical tool such as the Progress out of Poverty Index ® - that can be used to further segment customers into natural affinity groups. This will enable the organization to better understand how it is meeting its social mission, how poor customers respond to different product offerings, and make operational decisions based on the outcomes.
2. A clearly defined strategy and data architecture must be established at the outset. This should include leadership buy-in and support for the process; improved awareness of data privacy and legal issues with using and sharing the data (for example, to a third party like Grameen Foundation that may help with the analysis); and knowledgeable IT staff with database administrator rights who can retrieve the correct data from the system effectively in a way that also complies with and respects privacy laws and banking regulations.
3. If the ideal level of data is not available when you begin, take the first steps anyway and start analyzing the data you do have. There will be bugs to resolve in the data, but you will be able to do that through practice. Only then will you be able to learn something from what you have, draw some interesting insights, and hopefully use this information to garner the interest and investment in the additional infrastructure or resources required to obtain the additional data.
4. Data analytics should support the use of a management dashboard that will allow senior managers to view trends and evaluate design features to support their own decision making. This will also help to drive more interest and investment in data analytics. Similarly, weekly dashboards for product teams will be critical in evaluating product pilots and new marketing campaigns.

Annex: Building a Data Analytics Process

Grameen Foundation worked with CARD Bank to outline a comprehensive data analytics process that would leverage its data, understand its clients, and finally, implement solutions. Applying this process at your own institution can be done in a linear way, although we recommend restarting the process each time objectives shift. You should also keep in mind that analytics are inherently an iterative process as you discover data errors and the analytical process brings to light new questions.



Requirements

A conversation with management is required to set expectations and address concerns about any savings project. Senior management must agree on the goals of the project and have business questions that should be addressed. It is important for them to understand how a data analyst can help the institution answer those high level questions (such as “how can we improve our savings program?”) and reach those goals. One potential way to frame the discussion is to review the reporting needs to see what data is required for decision-making.

Staff resources should also be discussed. Mid level management staff need the required skill set to ensure that they can implement analytical work as part of their daily tasks. Information technology (IT) skills are essential as IT is the primary source of information, or data.

The role of the analyst is the second component in the equation. The required skill set for this type of work could be placed at a crossroads between social sciences, business and marketing, statistics and computer science. Thus many professionals start in one of those fields and build their expertise. A more experienced analyst will be well versed in trouble shooting, since data problems and other roadblocks are the rule rather than the exception in this type of work. Analysts must also be able to work across teams as they interact with many different actors including business and operations managers, database administrators, and marketing and sales teams

The data set required to meet that goal should be discussed with IT at the outset. The IT department must have the capacity to provide the data needed for the institution to begin the analytics process. Data inconsistencies are very common and almost always inevitable, so institutions should be properly equipped with the software and team skills needed to get the best possible data. Involving the IT team early will also help ensure there are consistent data structures available throughout the process.

Data Analytics Objectives

An institution must formulate an analytical objective that helps answer its original goal and business question. This is the transformation of a high-level goal into an actionable objective. Having an objective provides direction throughout the data analytics process, and guarantees valuable findings about clients. It is central to staying aligned with the institution's mission, such as reaching the poorest through new savings products or changing product to maximize its use by the poorest. The data analytics objective guides an institution to ask the right questions throughout the process, like “what type of products do the poorest use most?”

Selecting Appropriate Data

It is important that the data analyst and the IT department understand the goals and challenges of the institution. Some of the variables that can be utilized by pro-poor institutions on product performance and poverty are: PPI score, membership date, opening dates product balances, total loan amount disbursed, loan cycle, repayment rate, and demographics. Ideally, an institution should pull monthly snapshots of data from clients with a census or representative random sample. Further, data analysts must be cautious about disregarding variables since they may prove to be valuable when combined with other data.

The team should balance the availability of data with the effort required to get more data. The data that is available will shape the type of analysis. Keeping consistency throughout the process of pulling monthly snapshots of data is essential. If the query to pull the data changes over time, an institution will have different data structures; some variables may no longer be comparable or even exist in previous datasets.

Fixing Problems with the Data

Problems are very common with raw data, so the data analyst should perform a quality control check. This step is significant because an institution needs the most accurate data possible to create the best models and assumptions. If the data is not consistent, the models will not be dependable. To accomplish this, it is critical to establish good communication between IT and the analysts. As mentioned previously, it is important to communicate with IT staff about the type of data structure an institution needs, and suggest ways data problems can be prevented or remedied. For example, when data analysts find errors, they should attempt to fix it, go back to IT, or completely delete it to avoid skewed results. Some of the most common errors when performing data analytics are: out of range values, data redundancy, default values, and pollution.

Getting to Know the Data

Investigating data is in the discovery phase and the most creative part of the process. To determine which variables to compare, it is imperative for a data analyst to fully understand the unfamiliar dataset. By carefully exploring the raw data one can begin to understand the range of variables and note missing

values, outliers, and other inconsistencies. The analyst should comb through the data, noting the range of each of the characteristics, and paying particular attention to trends. For instance, a data analyst could pull the data from a particular characteristic, such as gender, and simply create a chart to give a visual representation of the gender distribution. The data analyst may also create a chart to illustrate the poverty levels of all clients. Data visualization reveals data that may be restricting the analysis. Data visualization includes creating histograms and scatter plots, as well as identifying outliers, out of range values, and missing and default values. Getting to know the data will also allow the analyst and the institution to segment the data.

Transforming Data

The next step involves transforming raw data into useful information. This involves converting the data so that it can display another level of measurement or combine several variables into one. An institution should build new variables by changing numeric variables into nominal ones, converting values into proportions, or capturing trends. By creating and combining variables, client findings come to the surface. Data becomes transformed into valuable findings about the institution's clients.

Building Models

The next step consists of creating models. This involves conceptualizing models to display how data relates to each other. Statistical software is necessary to use algorithms to crunch many variables or predictors in order to “explain” an outcome... Creating models is critical since the goal is to understand clients and the possible relationship between variables. This allows an institution to make rule based models and make predictions about future customer behavior. The model should represent a holistic view of the data by using information from the past to describe relationships between variables and ultimately predict about future outcomes. Since building models take a lot of time, effort, and resources from the institution and the analyst, it is vital to implement, test, and evaluate them. Finally, the best model does not necessarily have the highest level of statistical accuracy, but does have the ability to help reach the institution's goals.

Assessing and Deploying Models

It is critical to assess a model by testing its performance over time with new data. This will confirm that the model continues to be stable, accurate, and comprehensible. After implementation, the institution can assess its results. This indicates marketing to clients differently under certain conditions, or having a differentiated product offering, depending on specific needs. An institution should compare the results of applying the model versus not implementing it to ensure that there is a gained benefit from the model, rather than a setback. Further, it is necessary to continue testing the model every six months to a year. The results should be used to incorporate lessons learned for the institution and refine the model in the future.