REFLECTIONS ON METHODOLOGY

OCTOBER 2017

Using Big Data to Understand Borrowers of Subprime Loans

By *Kelsey Schaberg* and *Richard Hendra*

This post is one in a series highlighting MDRC's methodological work. Contributors discuss the refinement and practical use of research methods being employed across our organization.

The Subprime Lending Data Exploration Project is a "big data" project designed to produce policy-relevant insights using an administrative data set provided by Clarity Services, Inc. The data set, covering nearly 50 million individuals across the nation who have applied for or used subprime credit, contains information on borrower demographics, loan types and terms, account types and balances, and repayment histories. The first report reveals a wide diversity in borrower characteristics. This view of the "demand side" of small-dollar credit markets can inform the development of effective approaches to improve outcomes for users of high-interest debt.

In order to better understand how to help such borrowers, it is necessary first to understand what is driving usage. One of the key research questions of the study was whether there were distinct groups of borrowers in terms of loan usage patterns and outcomes. We hypothesized that there would be; the challenge was how to identify these segments of borrowers. In program evaluation, subgroups are often chosen based on individual participant characteristics that are expected to moderate the impacts of the program. But there has been very little research looking at segments of payday loan users on which we could base subgroup creation. To overcome this problem, we used a data discovery process called *K-means clustering* to identify segments of borrowers in the data.

K-means clustering is an "unsupervised" learning algorithm. In a "supervised" algorithm, such as regression analysis, the goal is to explain or predict a key dependent variable of interest. In unsupervised learning, there is no clear single dependent variable to organize the analysis; instead, patterns emerge based on multiple variables. K-means clustering places observations into groups based on their distance to the "centroid," or mean, of a group. The "K" in K-means refers to the number of groups — which is set by the researcher in advance — and the "means" refers to the algorithm placing observations into the group with the nearest mean. The basic goal of K-means clustering is to maximize the variation between clusters while minimizing the variation within clusters. This approach to clustering allows groups — or in our case, segments of borrowers — to emerge based purely on underlying similarities, rather than on researcher preconceptions. K-means clustering works in three main steps: variable selection, data preparation, and clustering. The sample used in the clustering analysis included 198,499 subprime loan users.

Variable selection. The first step in the identification of clusters was to understand the underlying dimensions of the data. One common problem in big data analysis is known as "the curse of dimensionality," in which the number of variables from which to choose in a data set is far greater than the number of true dimensions. For example, the data set had multiple variables measuring loan usage, one true dimension of the data set. High-dimensional data can undermine a cluster solution, particularly if dimensions are redundant.

In this project, we used a combination of researcher discretion and data-driven methods to reduce the variable set to the true underlying dimensions. Essentially, the data set contained variables on loan usage, loan volume, loan type, lender type, and borrower demographics. We decided to cluster based on the loan-related variables, using the demographic variables only later to see how demographics varied by cluster. Then we used an algorithm called VarClus, which identifies correlated variables using an application of a method called *principal components*, to help identify the

NEW YORK
16 East 34th Street
New York, NY 10016
Tel: 212 532 3200

OAKLAND
475 14th Street
Suite 750
Oakland, CA 94612
Tel: 510 663 6372

WASHINGTON, DC
1990 M Street, NW
Suite 340
Washington, DC 20036

LOS ANGELES
11965 Venice Boulevard
Suite 402
Los Angeles, CA 90066

www.mdrc.org

true dimensions of the data set and further narrow down the list of variables. The method works by identifying groups of variables in which the variables are as strongly correlated as possible with each other and as uncorrelated as possible with variables in other groups. After the variable groups (or dimensions of the data set) were identified, we selected one variable from each group to represent that group — based on researcher discretion and a measure of which variable was most correlated with other variables in its group and least correlated with variables in other groups.[1]

Data preparation. The K-means algorithm is notoriously sensitive to outliers and units of measurement. Variables that have larger variances exert more influence in determining the clusters than variables with smaller variances, and the validity of the statistics associated with evaluating cluster solutions can be threatened if outliers are included. To avoid having the results swayed by variables with high variances, the variables were standardized to z-scores, as is conventional, with a mean of 0 and a standard deviation of 1.[2]

Clustering. This step involves first choosing the number of clusters (or K) to include. We evaluated cluster solutions with 2 to 20 clusters in intervals of 2. This range was selected to allow the data to drive the solution. We wanted to identify at least two groups of borrowers, but we thought having more than 20 groups would result in clusters with sample sizes that were too small.[3]

Cluster solutions can be evaluated using a few different statistics, including the cubic clustering criterion (CCC) and the approximate expected overall R-squared. The CCC is a statistic developed by analysts at SAS, derived from the R-squared statistic. CCC values above 2 or 3 indicate a good cluster solution. In order to choose the number of clusters, analysts inspect a variety of plots — including plots of both the CCC and the R-squared by the number of clusters — that communicate the percentage of the variation in the data that is explained by the clusters. Analysts look for inflection points in these plots (places where there is no or only marginal improvement in the model) and balance the desire to explain more variation with the need to have an efficient number of clusters. If clusters are small or not easy to name, it will be hard for policymakers to take action based on the results. This process is explained in an appendix to the project report.

What did the cluster analysis reveal? Our final cluster solution had six clusters. We ultimately focused on three key clusters, which contained over 94 percent of the sample. These three clusters of borrowers differ greatly in the kinds of loans they use, the lenders from whom they borrow, and their loan outcomes. Borrowers in the largest cluster, constituting over 40 percent of the sample, struggle to repay loans and have the financial profile that has come to be associated with payday loans. However, there is another relatively large cluster of borrowers (roughly one-third) who pay back their loans on time and rarely default. A smaller, more distinctive cluster borrows mostly from tribal lenders — financial organizations affiliated with Native American tribes, which are generally not subject to state regulation. Compared with the other clusters, these borrowers fall somewhere in the middle in their default rates, are more likely to be in states where payday lending is restricted, and are more likely to use subprime installment loans rather than payday loans.

As more big data sources become available to researchers we expect to see increasing use of unsupervised learning techniques, such as K-means clustering, to provide more insight into the often heterogeneous participants in research studies.

[1] Variables with the lowest 1-R**2 ratio within their groups were chosen. This ratio is 1 minus the variable's correlation with its own cluster to 1 minus the variable's correlation with the next closest cluster.

[2] Standardizing variables to z-scores does not always take care of extreme outliers, which are a common source of small clusters. Outliers can either be deleted or have their influence reduced through transformation (such as top coding or binning into categorical variables). Outliers are a problem only during the actual cluster analysis, and the original variable values can be used for any further analyses of the clusters.

[3] The K-means clustering process was done with PROC FASTCLUS in SAS.