Anise Mulkey, Zach Brown, Kumar Raja Guvindan, Cory Nichols

[[1]](#footnote-1)

Anonymous Prediction: An Empirical Comparison

***Abstract* – To be completed**.

T

I. INTRODUCTION

HE value of personally identifiable information is undeniable. Marketers can tailor specific campaigns and product offerings based on a person’s income level, zip code, gender, preferences or a combination of features often available in public datasets today. These features can be married to other public datasets or used with statistical analysis methods to personally identify individuals with high confidence. Given an extensive set of features for a patient, medical information can easily be triangulated to identify a person, and subsequently, the person’s medical history. This information can lead to additional marketing or even worse: unwanted health status exposure to public entities, employers or hackers. Too often, datasets used for research contain features about individuals or entities that allow for the compromise of individual anonymity. In particular, healthcare research releases anonymized datasets used for the verification of research results. Combining these anonymized research datasets with publicly available datasets such as tax records or public membership lists could allow for the identification of an individual. Another practical example would be combining personal credit or loan payment records with the same publicly available datasets to identify individuals.

In this paper, we evaluate several methods to provide for and maintain the anonymity of individuals in a large dataset of personal loans. We also investigate whether anonymous data allows for accurate analysis results for a classification and a regression analysis task: predicting the loan outcome and loan interest rate. We compare

the results of our analysis based on anonymized data to the raw loan dataset.

The dataset used in our analysis is a publically available dataset of personal loans retrieved from Kaggle.com. The dataset contains 887,379 unique loans, with a mixture of 74 categorical and numeric features including features such as loan’s interest rate, home ownership status, employment length, annual income and the ultimate status of the loan.

Our analysis suggests there is a tradeoff between ensuring anonymity and obtaining meaningful analytical results. More specifically, we determine how much a dataset needs to be modified in order to be considered anonymized, while still allowing for meaningful analytical results.

Main results will go here.

Our paper is organized as follows: In section II, we give a brief overview of previous work and describe some of various anonymization techniques available for data today at a high level. In section III, we provide a detailed overview of our research and analysis methodology. In section IV, we proceed to exploratory data analysis and implementation methods of anonymization techniques and determine whether or not we are successful in truly anonymizing the loans dataset. In section V, we perform model selection, cross validation, and model comparison of the previously implemented anonymization techniques against a raw dataset. Finally, in section VI, we summarize our findings and address the limitations of our work. We provide recommendations for future efforts to address the limitations mentioned.

II. PREVIOUS WORK

Brute force methods such as data truncation or outright data exclusion have been utilized in the past to

achieve the anonymity of personally identifiable information in datasets containing people as objects or rows. Telephone numbers, addresses, social security numbers and credit card information are outright deleted from the dataset or truncated to a number of digits that represent an abstraction of the original data. However, the effects of brute force anonymization techniques, especially the deletion of features, can severely impact the results of data analysis. [1]

In a paper from the Office for National Statistics in the United Kingdom, administrative, personal data from healthcare sources was anonymized such that personal privacy was maintained but data could be re-used for census purposes. The goals of the research team were to find a methodology to provide more frequent census analysis capabilities while cutting costs of clerical matching and ensuring citizen privacy. Specifically, a cryptographic hash function was used to anonymize personally identifiable information retrieved from administrative datasets such as the National Health Register. These anonymous data were “linked” with existing census data to build a statistical representation of various counties in England and Wales. The research team used two methodologies: logistic regression and rule based matching on the anonymized dataset to confirm a match. Researchers measured the quality loss of their methodology by comparing against “gold standard,” clerical methodologies and performed only slightly worse in estimating actual census data for eight local authorities tested. [2]

III. RESEARCH AND ANALYSIS METHODOLOGY

The methodology and analysis for our research can be divided into three key parts: 1) pre-processing the data, 2) anonymization of personally identifiable characteristics and 3) the use of supervised machine learning techniques to a) predict the continuous outcome interest rate and b) predict group membership using a binary classification task and a discrete class label for ultimate loan performance disposition (good/bad). In the pre-processing phase, we address missing values in our data instances, delineate strategies for handling categorical data, extract the most meaningful features and eliminate the redundant ones (both loan-level and individual borrower features), conduct feature scaling where appropriate and partition our data into training and test sets. For the anonymization exercise, we provide details around the personally identifiable characteristics chosen and three standard methods utilized. Finally, we discuss the machine learning techniques that best interpret the potential conclusions (if any) from completing the experiment.

1. *Pre-processing the Data*

Our chosen Kaggle dataset is large and supervised machine learning techniques will undoubtedly be computationally expensive for a set of training samples this size. We narrowed our training set to a more manageable record count of 29,981 representing unsecured loans originated during an eight-year period from 2007 to 2015. This particular era is representative of wide variations in economic cycles, including the housing bubble bust of mid-2007, the subprime mortgage implosion and market correction in the following year and the ensuing recession and recovery. The macro-economic environment is a confounding variable in any loan performance evaluation and selection of origination vintages across this period of eight years should reduce the potential noise.

Detailed examination of the data revealed that all 74 attributes in the original dataset would not be necessary in our statistical quest for meaningful conclusions. Treatment of the attributes is summarized generally in the table below.

|  |  |
| --- | --- |
| **Count (#)** | **Attribute treatment** |
| *74* | *Original features/attributes* |
| 21 | Deleted: all records are blank for this attribute |
| 3 | Deleted: all records are the same for this attribute, not meaningful to the analysis |
| 5 | Deleted: not meaningful to the analysis |
| 18 | Deleted: redundancy, $$ amounts related to payments made or amounts invested OR strings, not material to the analysis |
| 47 | Total features deleted |
| *27* | *Total features in final analysis* |

In addition to eliminating both meaningless and redundant attributes, we modified four features to make them consumable for our machine learning algorithms. We changed “term” to “term\_mos” and removed the string “mos” from all records. Similarly, we also changed “emp\_length” to “emp\_length\_yrs” and removed the string “years” and the operator “>” from all records. We changed “earliest\_cr\_line” from a mm-dd-yyyy format to a continuous variable representing a number of months since the inception of a borrower’s earliest credit line. Finally, we changed loan\_status to a binary good/bad ultimate loan disposition indicator subject to the table below.

|  |  |
| --- | --- |
| **Binary Class Indicator** | **Original Loan Status Description** |
| Bad | Charged Off |
| Good | Current |
| Bad | Default |
| Bad | Does not meet the credit policy. Status:Charged Off |
| Good | Does not meet the credit policy. Status:Fully Paid |
| Good | Fully Paid |
| Good | In Grace Period |
| Good | Late (16-30 days) |
| Bad | Late (31-120 days) |

Categorical features, like loan purpose, were converted into boolean columns using Python’s one hot encoding functionality. We solidified feature selection from a list of 27 attributes via visualization-based comparisons of the default rates and the distributions of each borrower-specific feature. Our final features appear to represent the most significant differences in the bad and good populations of loans evaluated.

1. *Anonymization Techniques Employed*

After conducting extensive research into regulatory and business issues impacting the most widely-used anonymization techniques, final technique selections utilized in our analysis include k-anonymity, generalization and perturbation. These techniques involve data repetition, data grouping and data averaging, respectively.

In a data repetition or k-anonymity scenario, we artificially create additional records in order to “hide” sparse records. For instance, if the zip code 78729 occurs one time in our dataset, we create additional records containing 78729 as the zip code feature value.

Generalization of data is an anonymization technique that simply translates to removing the specificity from it by assigning values to more general categories. For example, the borrower’s annual income attribute is included in our data set. Generalizing the specific annual income values involves setting ranges for the data, i.e. “less than, $50,000, greater than $50,000 but less than $100,000, greater than $100,000 but less than $150,000.

Perturbation modifies the original data in a training set in a manner that is not statistically significant. There are several offshoots of the method, but we employ a practical “sub-technique” called microaggregation in our lending data. Microaggregation involves sorting the personally identifiable feature either in ascending or descending order, grouping similar-sized values, averaging those values and replacing the specific values with the appropriate mean.

Single methods, or combinations of methods to anonymize data are used to create a set of anonymized data sources. These modified sources are compared to a baseline, raw dataset containing unprotected personally identifiable information.

1. *Machine Learning Methods Utilized*

Our curiosity surrounding the impact of anonymization techniques on the validity of machine learning algorithm conclusions extends to two predictive scenarios: a continuous response variable interest rate and a binary ultimate loan disposition classified (good/bad).

Specifically, we predict the loan interest rate utilizing LASSO regression in order to make our feature set sparse and more efficiently capture important predictors. In general, we approach analysis from a machine-learning standpoint, utilizing algorithmic solutions such as LASSO for feature selection.

Obviously, anonymization techniques can affect the results of our data analysis. We compare predictive models objectively using the model root mean squared error (RMSE) on a stratified test data set partitioned off prior to model selection. A baseline model is fit on the full feature loans dataset with no anonymization. Subsequent models utilizing the same learner, LASSO regression, are fit on iterations of our “anonymized” loan dataset and compared utilizing the root mean squared error measure.

1. *Analysis and Results*

*Prediction of Continuous Outcome Interest Rate*

*Prediction of Group Membership*

*Key Takeaways and Inferences*

Below are a few key takeaways that may have affected several of the outcomes of the overall experiment:

REFERENCES

1. G. Nelson. (2015). “Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification.” Thotwave Technologies. Chapel Hill, NC. [Online]. Available: <http://support.sas.com/resources/papers/proceedings15/1884-2015.pdf>
2. D. Hsu. (2015, Apr.) “Techniques to Anonymize Human Data.” Datasift. United States. [Online]. Available: <http://blog.datasift.com/2015/04/09/techniques-to-anonymize-human-data/>
3. E. McAllister, T. Grance, K. Scarfone. (2010, Apr.). “Guide to Protecting the Confidentiality of Personally Identifiable Information (PII).” National Institute of Standards and Technology. Gaithersburg, MD., United States. [Online]. Available: <http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf>
4. W. Kan. (2016, Jun.). Lending Club Loan Data, 2016. Lending Club. United States. [Online]. Available: <https://www.kaggle.com/wendykan/lending-club-loan-data>
5. Office for National Statistics. (2013, Jul.). “Beyond 2011: Matching Anonymous Data.” United Kingdon. [Online]. Available: <http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/beyond-2011-matching-anonymous-data--m9-.pdf>

1. Anonymous Prediction: An Empirical Comparison is submitted for review by Dr. Daniel Engels on Sep. 6, 2016. This work was supported in part by the MSDS program at SMU.

   Anise Mulkey is a student in the SMU MS Data Science Program from Dallas, Texas (e-mail: amulkey@mail.smu.edu)

     Zach Brown  is a student in the SMU MS Data Science Program from Boston, Massachusetts (e-mail: [zbrown@mail.smu.edu](mailto:zbrown@mail.smu.edu))

     Kumar Raja Guvindan Raju is a student in the SMU MS Data Science Program from San Diego, California (e-mail: [kguvindanraj@mail.smu.edu](mailto:kguvindanraj@mail.smu.edu))

     Cory Nichols is a student in the SMU MS Data Science Program from Austin, Texas (e-mail: [pcnichols@smu.edu](mailto:pcnichols@smu.edu)) [↑](#footnote-ref-1)