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, census data or public membership lists could allow for the identification of an individual. A consumer driven example of unwanted personal identification involves personal credit or loan payment records. Combining publicly available datasets with “anonymized” credit history or loan payment data could easily be used to identify a person without their consent. Individual privacy, then, must be protected at all times, even for research purposes. However, the very nature of research places critical importance on accurate analysis results. The tradeoff of privacy and accurate analysis is a subject of extreme importance in the era of big data.

Our objective is to determine if anonymous data can produce meaningful analytical results. 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 classification and regression tasks. 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. 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 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. We compare predictive models using the previously implemented anonymization techniques. against predictive models using a raw, de-anonymized dataset. Finally, in section VI, we summarize our findings and address the limitations of our work. We also 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 if care is not taken to maintain the statistical relationships of features in a dataset.

In a paper from the Office for National Statistics in the United Kingdom, administrative data from a national healthcare source was anonymized such that personal privacy was maintained but data captured from the health care source could be re-used for census analysis 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 rules based matching on the anonymized dataset to confirm a match with census data, while protecting the true identities and health status information about each person in the dataset. Researchers measured the quality loss of their methodology by comparing against “gold standard,” clerical methodologies. The anonymous matching model performed only slightly worse in estimating actual census data for eight local authorities tested.

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 ultimate loan outcome as part of a binary classification task. 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*

We narrow our dataset 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 burst 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. Reducing the size of the data also provides computational benefits for cross-validation and testing.

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.

TABLE I

Attribute Treatment and Data Pre-Processing

|  |  |
| --- | --- |
| **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, unique identifiers |
| 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 transform four features to make them consumable for our machine learning algorithms. We change “term” to “term\_mos” and remove the string “mos” from all records. Similarly, we also change “emp\_length” to “emp\_length\_yrs” and remove the string “years” and the operator “>” from all records. The previous transformations allow us to represent features as integers. We change “earliest\_cr\_line” from a date format to a continuous variable representing a number of months since the inception of a borrower’s earliest credit line. Finally, we change loan\_status to a binary good/bad ultimate loan disposition indicator subject to the table below.

TABLE II

Loan Status Result: Binary Transformation

|  |  |
| --- | --- |
| **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 such loan purpose, were converted into Boolean features using one hot encoding. In total, 27 features are kept after visualization-based comparisons and detailed exploratory data analysis. Thus, our truncated feature set represent the most viable data points for our anonymization and analysis dual objective.

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.

Combinations of methods to anonymize data are used to create different anonymized datasets. These modified sources are compared to a baseline, raw dataset containing unprotected personally identifiable information utilizing ARX, open source software used in the medical industry to gauge the risk for personal identification in datasets. Acceptable anonymized datasets, based on ARX’s HIPAA scoring and scenario models are kept for further analysis.

1. *Machine Learning Methods Utilized*

Two classical machine learning tasks are performed after data anonymization to address the second part of our objective. A continuous response feature, interest rate, and a binary categorical feature, loan disposition, are used as target features for regression and classification tasks, respectively.

Records from the de-anonymized loan dataset are partitioned into training and test datasets for each machine learning task. Specifically, stratification is employed for splitting data for the classification task. Stratification ensures the distribution of classes in the test set is consistent with the distribution of classes in the entire loans dataset. Each anonymized dataset is also partitioned into training and test sets based on the results of training and test splits for the raw loan dataset. For instance, if record 25 falls into the test set after splitting the raw dataset, it will also be in the test set for each anonymous dataset. This allows for consistent, accurate comparisons of models based on each dataset.

For our regression task, we optimize model fits for raw and anonymous datasets utilizing LASSO regression. The LASSO algorithm allows for sparse feature sets and more efficient capture of important predictors by utilizing the L1-norm during optimization.

A logistic regression approach is taken for our classification task. Simple classification models are known to bias toward outcomes that occur more often. To address this tendency for overfitting and optimize model fits across raw and anonymous datasets, we fine-tune the performance of our models by adjusting regularization parameters and a penalty strength parameter.

For both machine learning tasks, we optimize models for each dataset (raw and anonymous) using ten-fold cross-validation and grid search. This method allows for a quick search of optimal parameters for each machine learning task and ensures randomness is controlled. Cross-validation is performed on the *training set* for model optimization to estimate generalization performance and tune parameters without bias. Each task requires unique measures for performance. For the classification task, the F1 score is used. F1 allows us to examine each classification model’s ability to minimize false positive and false negative outcomes. For our regression task, we compare results of each model objectively using the model root mean squared error. One optimized model fit is kept for each dataset based on cross-validation performance.

To truly answer our general objective question, we test each optimized model on the test dataset utilizing the same performance measures. Further, statistical tests are used to determine if model performance differences are significant. In particular, 95% confidence intervals are used to answer our general question: can anonymous produce meaningful analytical results? For reference and further investigation, we also compare each model’s final feature set and associated weightings.

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)