[[1]](#footnote-1)

Privacy and Accuracy: An Epic Struggle

***Abstract* –** Individual privacy is both an ethical issue and personal expectation. As a logical extension, the privacy of an individual’s personal data is both an ethical issue and expectation in today’s digital, big data world. On the other hand, the importance of data analysis and the discovery of new information has exponentially increased the need for access to and storage of massive amounts of data. The proliferation of data made publicly available today is a threat to individual privacy. Personally Identifiable Information (PII) can be acquired without authentication and used to identify individuals’ health or financial status, interests or even more private information such as a social security number. PII and the need for “more data” in general presents a supposed opportunity for better analysis in many cases. In this paper, we explore the impact anonymizing financial data has on the results of a classification model. We investigate the anonymity of a publicly available dataset of loans and compare the effectiveness of a data analysis method on the original loans dataset and an anonymized loans dataset. Results go here pending finalization of anonymization data.

I. INTRODUCTION

P

ERSONALLY identifiable information (PII) is both an asset and a liability. It can be part of a contract, taken illegally, and even given away for free knowingly or unknowingly. PII is a treasure for businesses, identity thieves and hackers alike. It can be used by itself or in combination with other information to identify an individual. Given enough data, such as a work title, address and age, or a unique identifier such as a social security number, a person’s identity can be discovered and used for business or more sinister purposes. 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. Given data for a patient,

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

medical information can easily be mined to identify a person and their associated medical history. If a dataset

lacks information used for identifying an individual, additional features can be married to the original dataset by using public datasets or performing statistical analysis methods to personally identify individuals. These possibilities threaten individual privacy and violate ethical frameworks for data privacy. An analyst need not

work in a confidentially-privileged role to obtain personally identifiable information. Public datasets used for analysis are readily available and present dangerous possibilities for identifying the people associated with data records.

Too often, public datasets contain features describing individuals or entities that allow for the compromise of individual anonymity. In particular, healthcare research releases supposed 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 has been shown to increase the possibilities of identifying an individual in the anonymized data set [1]. A consumer driven example of unwanted personal identification involves credit card metadata. Analysts were able to accurately re-identify credit card users in shopping malls using nothing other than their purchase histories, location, date and price of each transaction. Unique identifiers were used to label each customer. The researchers in this case did not need to use other data sources outside of the transactions and transaction metadata [2].

Individual privacy must be protected at all times, even when data analysis is the main objective. Specifically, individual privacy is protected when the data can be considered anonymous. This means no relationship can be established between a person’s identity, such as their name, and the data available. However, the very nature of data analysis places critical importance on accurate results. By anonymizing a dataset, the analyst risks losing critical information for their analysis. Therefore, a tradeoff between privacy of the data and accuracy of results is inherent when analyzing personally sensitive datasets.

In this paper, we de-anonymize records in a public dataset of personal loans by appending additional data available from public social media APIs. We analyze the original loans dataset and transform features with high risk using anonymization techniques to produce a safer, more private dataset. To address the previously mentioned privacy-accuracy tradeoff, we also perform a classification analysis task using both the modified loan dataset and the original loan data to predict the outcome of a loan. The results from each dataset are compared to determine if anonymization efforts have an impact on analysis results. Our objective is to analyze whether making the loans dataset safer impacts the accuracy of classifying the outcome of a loan.

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.

Main conclusions will go here after full anonymization is complete as part of final draft. Data still pending and debate still underway.

Our paper is organized as follows: In Section II, we give a brief overview of related work and describe various anonymization techniques and frameworks 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. We determine whether or not we are successful in truly anonymizing the loans dataset. We also explore the loans data and create two classification models: one based on the original dataset and a second based on a more private loans dataset with key PII features controlled. We compare these models to interpret the impact, if any, anonymizing key PII features has on data analysis. In Section V, we summarize our findings and address the limitations of our work. We also provide recommendations for future efforts to address the limitations mentioned.

II. RELATED WORK

A common example of data privacy standards and protection of PII comes from the Department of Health and Human Services (DHHS) in the United States. DHHS standardized the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule to protect health information. The Privacy Rule’s goal is to “assure individual’s health information is properly protected while allowing the flow of health information needed to provide health care.” It protects the privacy of people who seek and receive care. In particular, HIPAA protects “individually identifiable health information” and applies data anonymity standards to demographic information relating to past, present and future physical or mental health, the provision of health care and financial information. The Privacy Rule also sets standards for health data disclosure, authorization of usage and establishes severe penalties for violation of the Privacy Rule. [8]

PII associated with an individual’s finances is also a major privacy concern. Financial privacy is a baseline consumer expectation and an ethical issue associated with data in the financial industry. Privileged users may have access to financial PII as part of their jobs or a contract. However, in the public space, it is problematic if user identities can be mined from publicly available financial datasets. Therefore, public personal loans data should be protected, even if the data is to be used for analysis or research. An analyst, hacker or identity thief could easily violate privacy expectations of consumers and make an assessment on a person’s credit-worthiness, or worse, impersonate the person on other financial applications for loans if given enough data. In the realm of data analysis, however, additional data is an asset. Data privacy may actually impede analysis, creating the tradeoff previously mentioned between privacy and meaningful results.

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. [8] 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 [7] 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.

A famous example of combining data to identify users comes from the Netflix challenge. The Netflix Challenge was a competition held to provide Netflix with a better recommendation system for movies. While the data used for the competition was considered anonymous by Netflix, user names were stripped out for instance, analysts were able to combine internet movide database (IMDB) data, including user names and profile information with the “anonymous” Netflix data. This allowed the analysts to identify many of the users in the Netflix dataset. This information allowed the researchers to determine a user’s political orientation, religious views and even body type based on movie reviews. [10]

III. ANALYSIS METHODS

Our objective is to determine whether making the loans dataset private impacts the accuracy of classifying the outcome of a loan. The methodology and analysis for our research can be divided into three key tasks: risk assessment and anonymization of personally identifiable features, pre-processing of the loans dataset and the use of a supervised machine learning technique to predict ultimate loan outcome as part of a binary classification task. For the anonymization task, we empirically investigate the risk of the loans dataset and provide investigate PII features and features that can be transformed to identify an individual. In the data pre-processing phase, we address missing values in our data instances, delineate strategies for handling categorical data, extract the most meaningful features and eliminate redundancy (both loan-level and individual borrower features), conduct feature scaling where appropriate and partition our data into training and test sets. Finally, we discuss the machine learning techniques that best interpret the potential conclusions (if any) from completing the experiment.

1. *Risk Assessment and Anonymization Techniques Employed*

An empirical approach is taken to assess the identification risk of the loans dataset by appending social media data to the original dataset. This approach is a simple, straightforward method for determining identification risk associated with the loans dataset. We utilize Facebook’s Graph API and a linear rules based classification algorithm to match users with loan records. A recent sample of loans is taken and ran through the identification algorithm based on Facebook’s API to categorize users into loan records based on a simple confidence rating.

After conducting 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 a second, anonymized dataset. This modified dataset is compared to the original loans 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. *Pre-processing the Data*

For data analysis purposes, we narrow our dataset to a more manageable record count of 42,535, representing unsecured loans originated during a five-year period from 2007 to 2011. 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 reduces the impact of the macro-economic swings seen each year.

Detailed examination of the data reveals that all 74 attributes in the original dataset are not necessary in our statistical quest for meaningful conclusions. Several columns are entirely blank for every record, there is extensive duplication of information in our feature columns and some of the data is not meaningful to our analysis. We perform the following pre-processing steps to convert critical feature columns into consumable data for our machine learning algorithms:

* Changed term to continuous and deleted the string 'mos'
* Changed emp\_length to emp\_length\_yrs and deleted the string 'yrs', imputed 5062 <1 year values to 1 year
* Changed earliest\_cr\_line to continuous number of months since earliest credit line, using today minus earliest\_cr\_line date

Treatment of the remaining 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 |
| 4 | Deleted: all records are the same for this attribute, not meaningful to the analysis |
| 7 | Deleted: not meaningful to the analysis |
| 16 | Deleted: redundancy, typically dollar amounts related to payments made or amounts invested OR strings and not material to the analysis |
| 3 | Too much missing data, imputation not realistic |
| 51 | Total features deleted |
| *23* | *Total features in final analysis* |

Additionally, missing values must be mitigated prior to construction of accurate machine learning models. There are 29 records with blank values in multiple feature columns (delinq\_2yrs, inq\_last\_6mths, open\_acc, pub\_rec, total\_acc). 26 of the 29 records have a loan status of "Does not meet the credit policy: Fully Paid"; the remaining three had a loan status of "Does not meet the credit policy: Charged off. Given that our data set is unbalanced in favor of our non-default outcome and contains more than 42,000 records, we do not believe that dropping these 29 rows from consideration will materially impact the model.

During our exploratory data analysis (EDA), we discover that revol\_util, or the amount of credit the borrower is using relative to all available revolving credit, is impactful to our model. Instead of deleting the records that are blank for this important feature column, we decide to impute the missing values with the mean for revol\_util.

Emp\_length\_yrs has a material number of missing values and is initially interpreted as an object in Python. Imputation of values requires conversion of the string "n/a" into np.nan and conversion from an object to a float to prepare the feature for analysis. We imputed missing values for the emp\_length\_yrs attribute by grouping grade and calculating the median for the respective grade groupings.

The sub\_grade field is ordinal and requires conversion to an integer for appropriate consumption by our models. We would expect the grade and sub\_grade features to be highly correlated since the sub\_grade is equivalent to the grade with a deeper level of granularity. We should choose only one of these fields for our model and sub\_grade appears to have a slightly stronger correlation with the response than grade does. Therefore, we will drop the grade column and use sub\_grade in our model construction.

Annual income has a material number of outliers and an obvious right skew. To proactively address the issues this will introduce into our analysis, we log transformed annual income and present the outcome of that modification below. The distribution is much more appropriate for linear analysis now.

*Figure 2.1: Distribution of Log Transformed Annual Income*



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, home ownership and income verification status were converted into Boolean features using one hot encoding. In total, 23 features are kept after visualization-based comparisons and detailed exploratory data analysis. Thus, our truncated feature set represents the most viable data points for our anonymization and analysis dual objective.

1. *Machine Learning Method Utilized*

A classical machine learning method, logistic regression, is utilized to perform loan outcome prediction after data anonymization. Logistic regression is a linear learner utilizing coefficients or weights for features and a link function called the sigmoid to transform continuous responses based on those weights to probabilities for each loan outcome class. Using fitted weights and feature values, we predict a binary categorical feature, loan disposition, for both the anonymized loans dataset and the original dataset. We measure the success of prediction using accuracy and F1 metrics, which gauge how well our model predicts overall and how likely our model is to predict the correct outcome, respectively.

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.

IV. ANALYSIS RESULTS

1. *Anonymity of the Data*

**Note:** Full anonymity data coming in final draft. ARX methodology is still being considered for inclusion due to relevance/ease of interpretability. We believe our simple method below can illustrate a lack of privacy effectively. Although FB has, as of Oct 30th, removed even more functionality to prevent such behavior in their Graph API.

We take an empirical approach to gauge the risk of the loans dataset. The loans dataset contains PII features such as employee title, work experience in years, and zip code which are established PII data [11]. Some of these features are transformed to protect loan applicants’ privacy. However, these transformations are not adequate and maintain risk in the dataset. We utilize Facebook’s Graph API [14] to test the anonymity of a sample of the original loans data using an algorithmic approach. Of 75 loan applicants sampled, 9 individuals are uniquely identified. This method does not require in-depth statistical analysis, rather, feature scoring is used to estimate the confidence in a match between the loans dataset and Facebook profile data.

Five key features are used in our query to match loans to Facebook user profiles: Age range in years, location, job title, work experience in years and loan create date. Other obvious and viable PII data that can be used in chaining data sources together available in the loans dataset is income and home ownership status. These features were not used in our Facebook queries, as income and home ownership statuses are not standard fields available in user profiles. However, they would prove viable for chaining other data sources together, such as property tax and census data.

Zip code in the loans dataset is partially hidden for privacy reasons. The first three numbers, however, are available and can be used to identify locales of users applying for loans. We utilize publicly available zip code data to capture a loan applicant’s city via the first three numbers and use the state already available in the loans dataset to complete a full city, state location. Further, job title is particularly risky, especially for unique job positions at local businesses. Work experience allows us to infer the age of the user, or at least a range for age. Data such as location, recent activity and usage rates have been used previously to identify users by coupling Twitter data [13].

These fields can be searched algorithmically with Facebook’s Graph API and a rules based classification approach. API search queries in Facebook have been diminished over the past year due to privacy concerns. A robust SQL-like interface is now removed in favor of a stripped down wildcard search method with more limited capability. However, even with diminished search capabilities, a wildcard search is still capable of producing a list of user unique IDs. These unique IDs are then query-able and return fields we wish to match on to identify the user. Users are identified as potential matches based on a simple linear weighted scoring method such as

.

(1)

Match Score has a range of 0 to 5. Because of the specificity job title provides, a higher weight of 2 is used in the linear rules based model. A job title match is attempted from the loans data set to a list of employment history from the user profile on Facebook. If the entirety of the title is contained in employment history, the match is considered 100%, thus the JobTitle parameter is scored a 2 in the equation. Partial matches are also allowed based on counts of matching keywords in a bag of words for each record. Location is binary and is matched based on city, state in the loans dataset and the Facebook profile.

Outlining the lack of security in the loans dataset further, a loan applicant’s age can be inferred from the work experience listed on each loan application. An individual with eight years of work experience, for example, has an age range of 24 to 28. If a user falls into this category, it is scored highly (100% match). Secondary ranges are also used, with lower scores, to allow for fringe matches. For instance, the age range 23 to 29 is a secondary criterion, if matched, would result in a 75% match, adjusting the overall Match Score downward but not to zero. Similar logic is also applied to loan creation date in order to take into account the recency of the loan.

A random sample of 75 recent loans (two years or younger) were classified utilizing the previously mentioned algorithm. We were able to uniquely match 9 Facebook profiles with user names and minimize an additional 17 loans to five users or less utilizing a simple algorithm. Additionally, data including user group membership, personal interests and other personalized information from Facebook can easily be appended to the loans dataset. While accuracy of classification is not our objective in this example, we show that the loans dataset, a public dataset, is unsafe. As a note, Facebook has deprecated unique ID look ups as part of the Graph API as of October 30th, 2016 and further limited the ability to capture user identities via their API. However, even a simple search on Facebook’s UI would allow user identification. We believe this still provides substantial risk when coupled with other PII data.

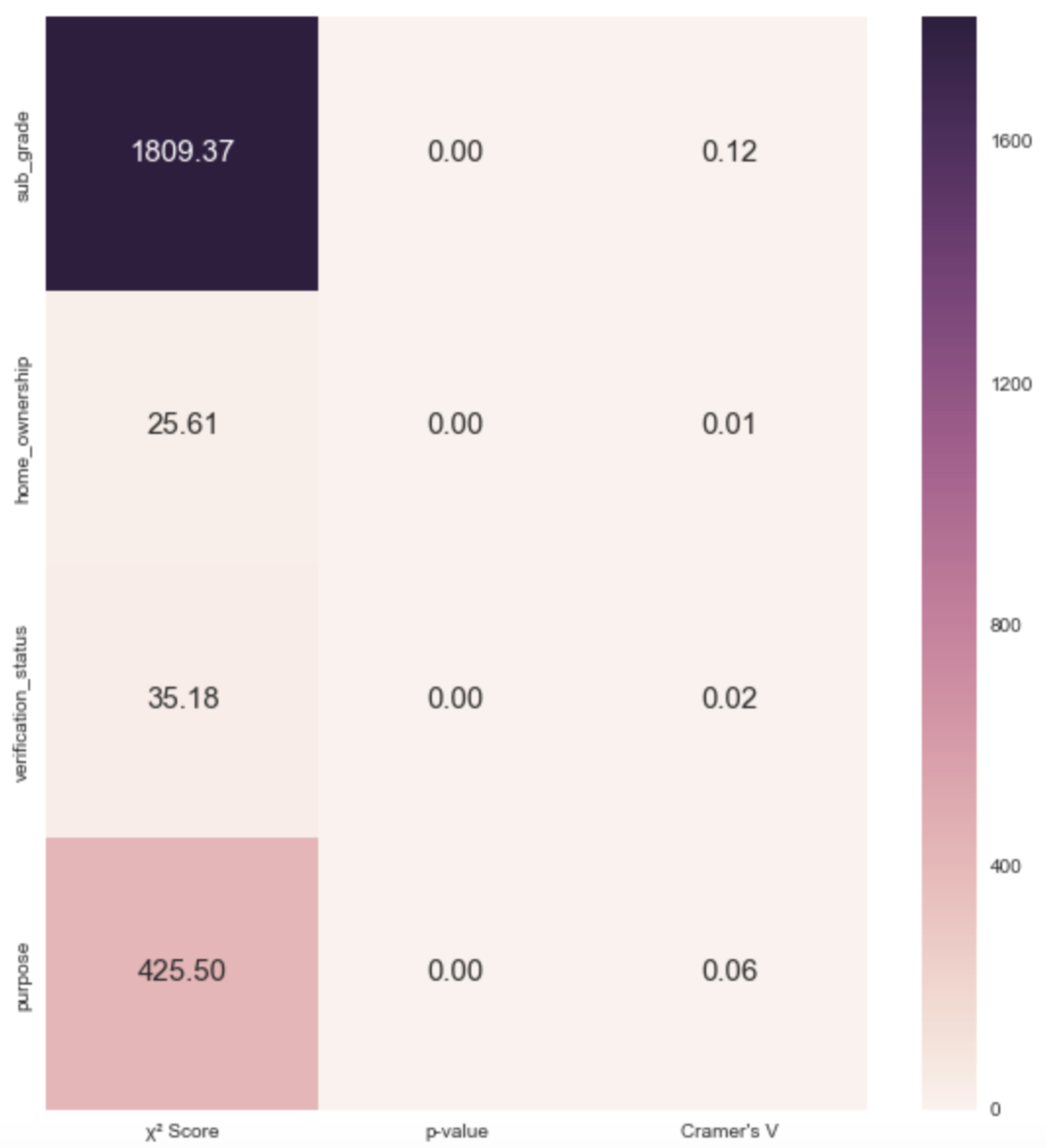
1. *Anonymization of PII Features Previously Explored*

Results from PII transformation go here as part of final draft. Not all data is ready due to healthy debate of anonymization techniques between group members and usage of ARX is still being considered.

1. *Data Pre-Processing and Prediction of Group Membership on the Original Loans Dataset*

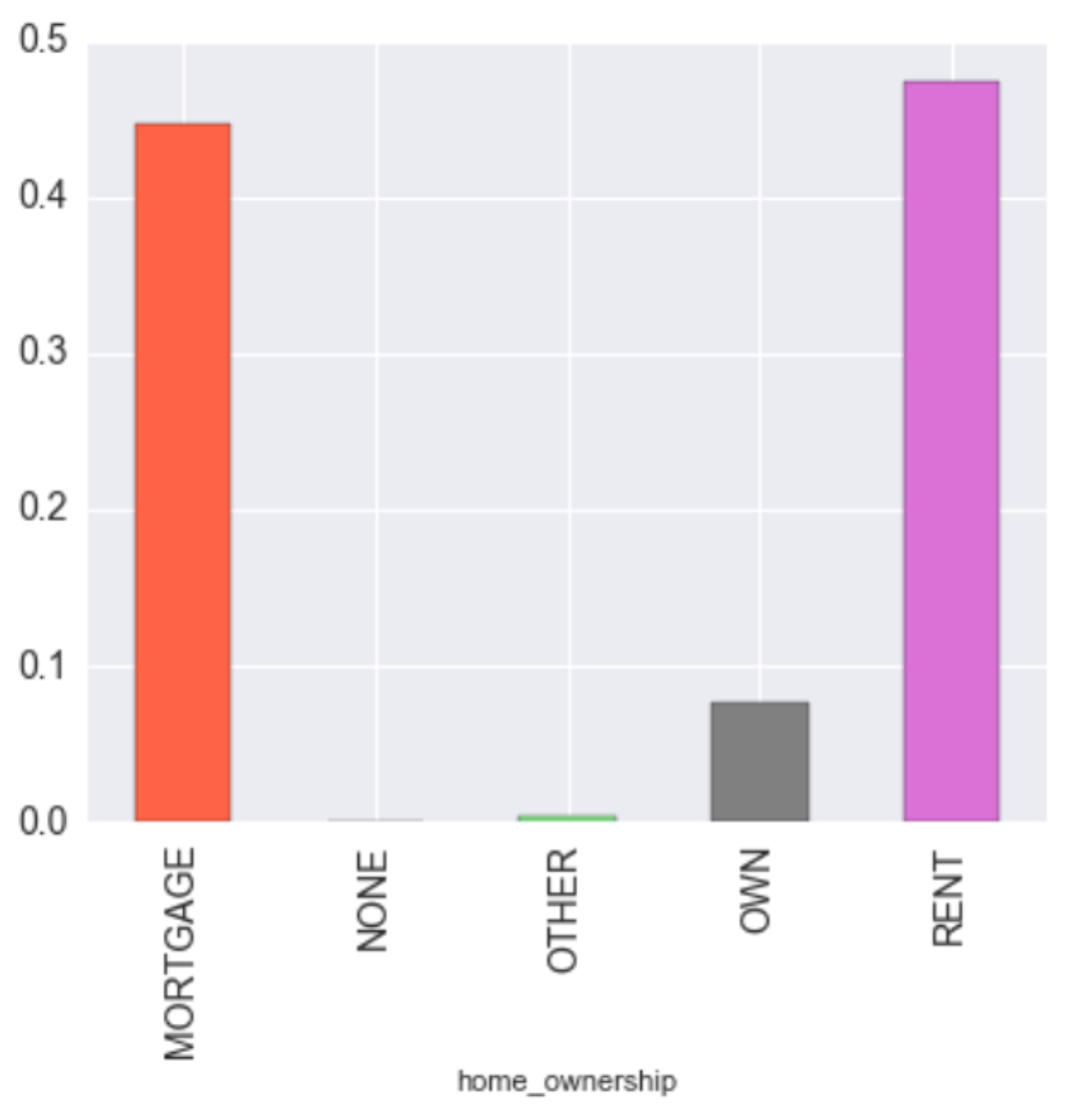
A chi-squared test of independence on our categorical features shows that all are significantly related to our dependent variable Default. The most dependent variable is sub-grade, which supports that ultimate disposition of the loan is highly correlated with the proprietary Lending Club grade and makes intuitivesense. Effect size, or strength of association, is also strongest (but considered moderately strong) between sub-grade and default at 0.12, which is only a "small" to "medium" effect. Income verification\_status has almost no association with the default disposition. Of the categorical features, sub-grade (upper left corner) would be our best predictor.

*Figure 2.2: Chi-squared test of independence categorical features*

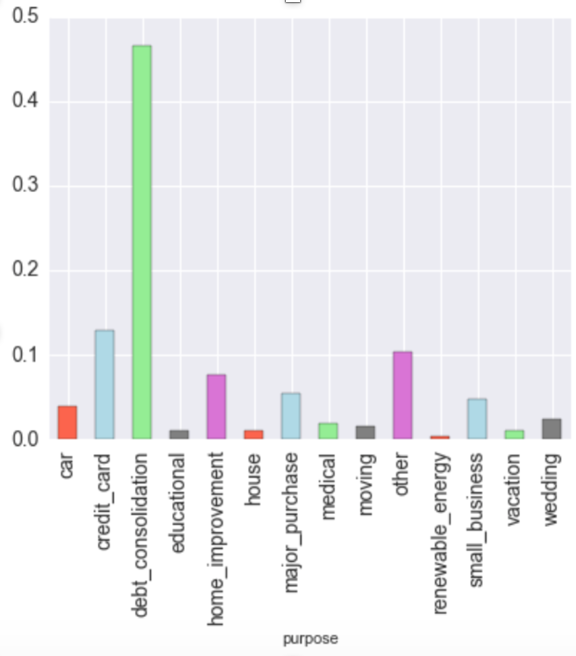


Examination of our categorical variables reveals several interesting points to keep in mind prior to buildout of our models. Home ownership status is fairly evenly weighted between borrowers with a mortgage and borrowers who rent as seen in Figure 2.3. Income verification seems like it would be very important in the extension of unsecured credit, but Lending Club did not verify income for almost 45% of the loans in this data set. We also determined that debt consolidation is by far the most common borrower purpose for Lending Club loans, seen in Figure 2.4.

*Figure 2.3: Home Ownership Status of Lending Club Borrowers*



*Figure 2.4: Loan Purpose for Lending Club Borrowers 2007-2011*



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 training and testing purposes, we separated 80% of the instances for training and the remaining 20% for testing.

We optimize models for each dataset (raw and anonymous) using ten-fold cross-validation. Cross-validation is performed on the *training set* for model optimization to estimate generalization performance and tune parameters without bias. We repeated this process of separating the testing and training data ten times. We utilize the holdout cross validation method built into scikit-learn.

To improve our model generalization performance and to determine the most optimal classifier hyper-parameters, we implement grid search cross validation. The GridSearchCV has a scoring parameter that is used to select the best parameter combinations.

We impose constraints upon the model to control overfitting via regularization, utilizing both L1 (forces most feature weights to zero) and L2 (penalizes large individual weights) methods. Assigning class weight equal to None, the average accuracy of the ten iterations using both L1 and L2 regularizations is almost equal at 85%.

Each task requires unique measures for performance. For the classification task, the default grid score uses accuracy score, which we quickly determine is not necessarily a good performance measure for our model. Classification performance will instead be measured using f1-score and the algorithm that has smaller false negatives will be used as best model. F-score allows us to examine each classification model’s ability to minimize false positive and false negative outcomes.

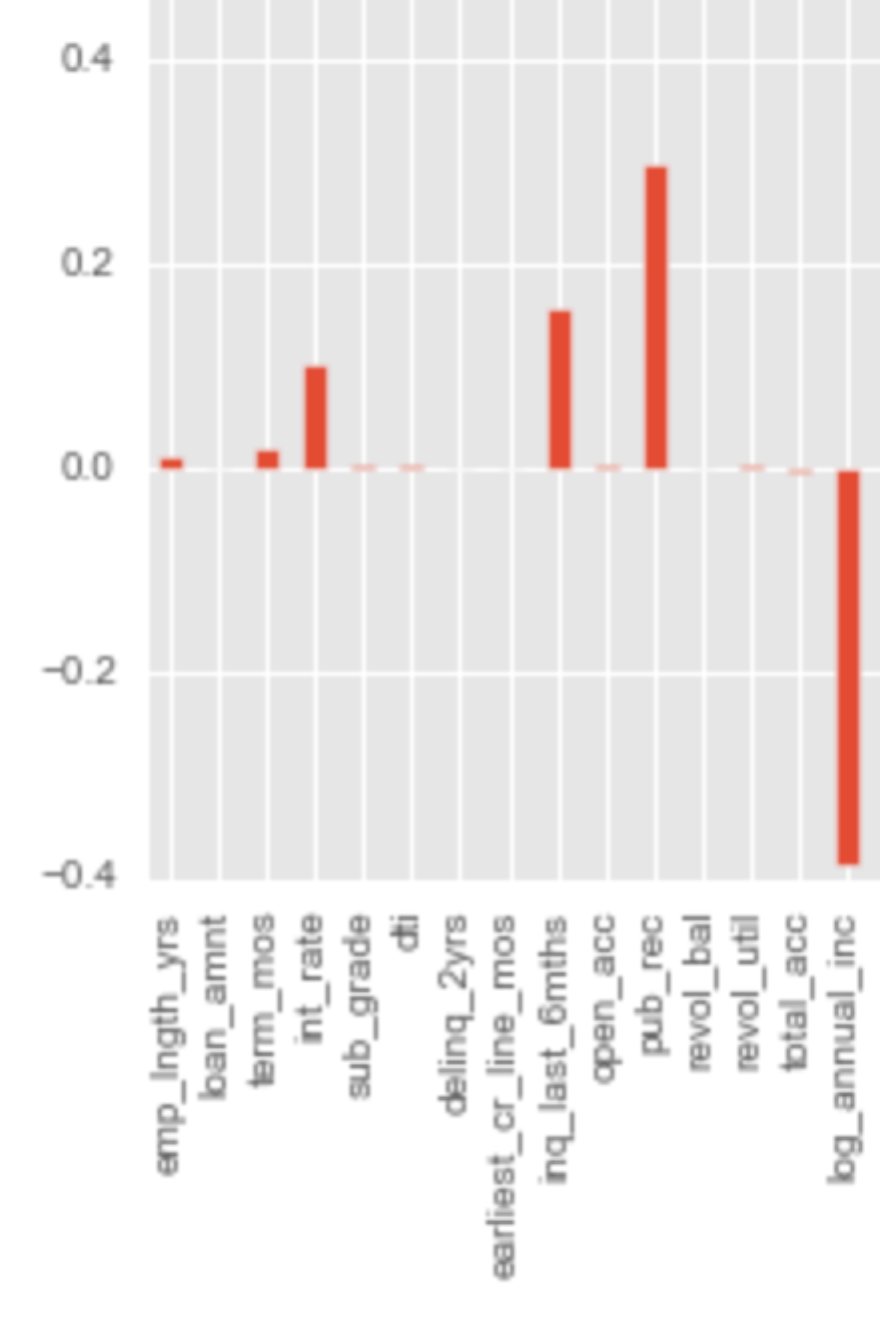
One optimized model fit is kept for each dataset based on cross-validation performance. Even though both L1 and L2 regularization produce specificity of 100%, sensitivity scores are negligible (0.01% for L1 and 0% for L2). The average precision scores of both L1 and L2 are also far below optimum. We measure our model performance based on how well it predicts the 'default' class. The current model is useless for such a purpose. This calls for attempting a better model.

We design and implement a strategy for managing the accuracy paradox in our data, i.e. the imbalance in our binary output variable (15% one and 85% zero). Any measures of model accuracy will reflect our underlying class distribution and must be weighted/adjusted accordingly. We increase the weight of class 1 relative to class 0 to mitigate the impact of this potential issue.

Considering the class imbalance in the output variable produces significantly different model predictions. The average accuracy and specificity scores reduced to 65% for both L1 and L2 regularization. On the contrary, the average sensitivity score increased to 65% for both regularization techniques. The precision score has not changed. We did not observe a significant difference in these scores between L1 and L2 forms. Moreover, there is no significant difference in the scores among the various iterations. Even though the sensitivity score has increased, the models still have very low precision. As a result, these models are far from optimum. Variances in cost variable and see if also fail to produce model improvements.

We interpret class weights using the coefficient of annual income as an example. Since the coefficient of annual income is negative, we can say that an increase in annual income is associated with a decrease in the odds of loan default. Conversely, the coefficient of public records is positive and the most significant of our continuous variables. This means that an increase in the number of public records is associated with an increase in the odds of loan default. Both of these observations are intuitive.

*Figure 2.5: Logistic Regression Coefficients*



We test each optimized model on the test dataset utilizing the same performance measures to address our objective.

As seen in the logistic regression coefficient breakdown, numerous PII features such as annual income, employment length, and home ownership are present and contain significant coefficient values. These features, especially income, play an important role in our analysis task. However, these features also allow for identification of loan applicants given the right set of companion data, as seen previously.

1. *Prediction of Group Membership on the Private Loans Dataset*

Note: Part of final draft, after anonymization techniques identified and fully vetted via ARX or another method still under heavy debate.

1. *Model Comparisons: Anonymization’s Impact to Data Analysis*

Note: TBD as part of final, after anonymization techniques identified.

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 records  
 produce meaningful analytical results? For reference and further investigation, we also compare each model’s final feature set and associated weightings.

V. FINAL OBSERVATIONS

Note: Final observations will be part of the final draft after anonymization techniques are identified.

REFERENCES

1. Lande, W., Rao, R. (2003, Jan.). “Secure De-identification and Re-identification.” Computer Aided Diagnosis & Therapy Group, Siemens Medical Solutions. Malvern, PA., United States. [Online]. Available:

https://www.media.mit.edu/research/highlights/unique-shopping-mall-reidentifiability-credit-card-metadata

1. Montjoye, Y. (2015, Jan.). “Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata” MIT Media Lab. Cambridge, MA., United States. [Online]. Available:

https://www.media.mit.edu/research/highlights/unique-shopping-mall-reidentifiability-credit-card-metadata

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 Kingdom. [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>
6. No Author. (2014, Mar.). “Overview of Data Anonymization.” [Online]. Available:

<http://www.odbms.org/wp-content/uploads/2014/03/The-Complete-Book-of-Data-Anonymization_Chap_1.pdf>

1. “Summary of the HIPAA Privacy Rule.” U.S. Department of Health and Human Services. United States. [Online]. Available:

<https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/>

1. Combining Datasets Add More Detail to Reference <https://www.schneier.com/blog/archives/2007/12/anonymity_and_t_2.html>
2. “Rules and Policies – Protecting PII – Priacy Act.” U.S. General Services Administration. United States. [Online]. http://www.gsa.gov/portal/content/104256
3. Federal Governmenet Zip Codes. [Online].

Available: <http://federalgovernmentzipcodes.us/>

1. Ghose, T. (2013, Apr.) “Think You’re Anonymous? Cell Phone Data Gives You Away.” [Online]. Available: <http://www.livescience.com/28353-anonymous-phone-data-not-anonymous.html>
2. Facebook Graph API. Facebook Inc. [Online]. Available: https://developers.facebook.com/docs/graph-api

1. Privacy and Accuracy: An Epic Struggle 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)