Developing a Data Science Approach to Detecting Income Fraud for the Peer to Peer Loan Industry.

David Keough, Nick Enko, and Brian M. Shake
College of Computing and Technology, One University Park Drive, Nashville, TN 37204

MSDS 5053 Practicum 1

Abstract - Personal loans can be obtained by borrowers from very different types of lending institution. The most common are a traditional loan institutions, payday lenders, or a Peer to Peer (P2P) lending brokers. P2P lending companies do not loan the money directly. They link the borrower to a lender and provide the lender with the borrower's income which is not usually verified. The P2P lenders collects fees based on the transaction and financially benefits from a higher number of introductions of borrowers and lenders. P2P lending is becoming more popular among borrowers because the pay highest interest rates are much lower than payday lenders and loans require less verification of income and assets than traditional loan institutions. A higher reported income with P2P lenders can result in a larger loan for the borrower and thus more profit and fees generated for the P2P lender. If the loan defaults due to an overstated or fraudulently reported income by the borrow, the P2P lender does not suffer, it is the lender that was matched to the borrower by the P2P lender that will incur the financial loss. This paper focuses on proposing a data science approach to detecting loan applicants that provide fraudulent income data to P2P lenders. The data obtained for this study contained 887,379 observations and 74 variables of loan applicants from the P2P loan company, Lending Club. The initial observations of this data showed that unverified loans make up 23% of the defaulted loans while verified and source verified loans made up about 77%. Described within this paper is how the data set for analysis was obtained and prepared for analysis, the initial findings, the proposed data science approach to fully analyzing this data, and the significance of the lending industry, both traditional and P2P, with a method of detecting fraudulent income reported on loan applications. Models generated from this analysis could be incorporated by lenders into their applications along with research in this area should improve the P2P lending industry by increasing the detection of fraudulent income reported on loan applications.

Keywords **Detect, Financial, Fraud Detection, Housing Crisis, Kaggle.com, Lending, Linear Regression, Peer to Peer Lending, P2P Lending, Predicting Income, Self Reporting Income, Unverified Loan, Loan Fraud**

I. INTRODUCTION

A. Problem Space

The problem being addressed is that of individuals reporting inflated income amounts on loan applications to qualify for higher loan amounts and secure lower interest rates. The effect of this fraud can lead to loan defaults which can touch every corner of the economy (Karlan,2008). It can generate a unhealthy lending cycle. When leander's margin and investment returns diminished from defaulted loans, this reduces the availability of funds for lending. Then only the more qualified and less risky borrowers. This reduces the spending power of the population, reduces the amount of goods and services that can be consumed, and further reduces the funds available to borrow and encourages more fraudulent loan applications. The earlier in this lending cycle that this can be detected the greater the benefit to the entire system of lending.

A higher reported income by the applicant will result in a larger loan amount due to the perceived ability to pay the loan back with interest in a relatively short amount of time. This can result in a loan for a higher amount with greater risk of default. Income fraud on loan applications were part of the problem that led to the housing crisis of 2008 (Adelino, 2015). During this crisis loans were approved by traditional leading institutions where the income reported was either not accurate or a result of fraud. The result of these indiscretions damaged financial institutions and the global economy. The lenders were already overstepping the traditional thresholds for mortgage lending practices and these additional deviations tipped the scales to allow a higher rate of default. Freddie Mac and Fannie Mae, two of the biggest mortgage companies in the country were bailed out with a cost of 187.5 billion dollars (Light,2015). The total cost of the bailout was approximately 700 billion dollars (Adelino,2015). Later in 2015 a new cost was reported to an order of magnitude larger at 16.8 trillion dollars (Collins, 2015). This is the largest bailout in the history of the United States, being hailed as too big to fail. Between 2006 and 2014 more than 9.3 million homeowners were either: foreclosed on, surrendered their home to a lender or sold their home via a distress sale (Kusisto, 2015).

The P2P lending industry has experienced significant growth after the housing crisis and recent reports indicate the industry has a higher percent of loan applicants with non-verified income than traditional lenders. While the housing crisis may have dimmed from the public eye, the behavior of reporting fraudulent income still continues, although in new forms, and its damage is still present.

On new form of lending is P2P lending. This is the practice of lending money to individuals or businesses through an online brokers service that match lenders directly with borrowers. These brokers offer a different range of services that range from: marketing, online transaction platform, models for approvals and pricing, borrower verification, credit checks, loan servicing, collection efforts, with legal and compliance assistance (Herzenstein, 2011). This type of lending differs from traditional loans offered by brick and mortar financial institutions. P2P lending companies do not loan the money directly. They link the borrower to a lender and provide the lender with the borrower's income which is not usually verified. P2P lending has become popular among borrowers because the pay highest interest rates are much lower than payday lenders and loans require less verification of income and assets than traditional loan institutions. One of the larger P2P lending institutions is Lending Club.

The specific P2P lender in this study is Lending Club. Lending Club has submitted Loan Data to Kaggle.com (Narayanan, 2011) which is publically available for analysis. Models generated from an analysis of this Lending Club data could be incorporated by lenders into their applications policies to help determine if the applicant's income is being misrepresented and protect the risk profile it is actually taking.

B. Motivation

One question that presses the review of the P2P marketplace is the motivation and the need of the market to exist in the first place. Without drivers, the emergence of the new capital exchange would not have been conceived. The issues that are brought to light are; what are the sources of the funding and what is driving its demand?

The largest use of a debt instrument in the United States is for the purchase of a single family residence. However, an increasing number of home purchases have been made not by buyers intending to live in the home, but by investors seeking a higher return on their investment (Kearns, 2016). In the past, occupant home buyers would contribute a certain amount of savings and secure the shortest term loan. The primary goal was to purchase a home and extinguish the debt as quickly as possible. As the derivative market began to further develop it created many diverse variables to allow for a greater variety to the home owners. "Credit derivatives are over-the-counter financial contracts that have payoffs contingent on changes in the credit quality of a specified firm or firms; the specified firm is

typically not a party to the contract." (Duffee, 2001) It allowed the term of the loan to extend from a range of a 10 to 15 year term to a range of 25 to 30 year term and even further. It created the balloon mortgage that allowed the buyer to pay a front loaded interest model that started with a lower payment and the payment would grow with the maturity of the loan. Interest only also became a model that allowed for a lower payment option and no real expectation of home ownership. This is a model that allowed for a family to occupy a single family residence as opposed to an apartment or multi-family situation. Green, R. K., & Wachter, S. M. (2005)

These niche options for capital acquisition were simply toggling between, risk, savings, likeliness and ability to repay the funds. The federal system also needed to accommodate the evolving landscape of funds. Pressure to own a home and become a part of the "American Dream" helped open the flood gate of federal funding available to lenders and potential buyers. This increased the access to homeownership while a blind eye was turned to the traditional mortgage model. Private Mortgage Insurance (PMI) was introduced to compensate for the lack of traditional savings to make the typical down-payment. The mortgage backed market was eager to provide a higher return as opposed to the sagging saving and certificates of deposit rates that were returning sub market returns for parked funds. Investors were seduced by the higher fund rates promised by the riskier portfolios that typically were a sound source of income in the more vetted and vested 80% loan to value 15 year platforms. Holt, J. (2009)

This lucrative opportunity created a two fold conundrum. The funds that were traditionally captive at the banks paying the low annual funds to investors were now going to the higher risk portfolios to drive the greater rate of return. The demand for funds had not been satisfied but simply redirected. The new variable derivatives were allowing cash out loans and driving the loan to value numbers to 90, 95 and even 100 percent of the value of the property. The second problem created was in the need to sustain an exchange for the marketplace. A marketplace has to possess both a seller and a buyer to sustain itself. According to Garmaise, M. J. (2015), the higher loan to value (LTV) and the cash out options had been redirecting funds into the purchase of personal goods and services and not solely for the purchase of the property.

The banks were no longer the primary provider for the mortgage products and were operating in a cash poor model as the availability of funds decreased. This forced an inflated rate to bring in investors. An increase in the cost of the funds followed, making it more difficult for individuals to borrow money for personal expenditures. The end of this string of events witnessed the arrival of the liquidity derivatives.

The question comes how do you allow a public to operate and participate in a consumption model when the majority of their number one source of wealth, income, is used up in servicing existing debt. We are a consumer's market. Less and less of the goods we consume come from a local market and as the purchase funds are driven off shore to quench our appetites, we develop primary and secondary markets to more goods but have no real reinforcement of our GDP and we begin to see the wheels of consumption slow.

Potential home buyers have an increased interest in the P2P market as a solution to an environment created when funds become less available to them when they want to buy a home. Addressing the limited availability of funds for purchasing a home is a growing problem requiring a solution. The constant exchanging and servicing of existing debt, increases the income and debt ratio gap. Consumers faced with this situation are under an ever increasing burden. This is the situation where the likelihood of consumers misreporting their income increases. In effect this increase is a response to the consumption bubble.

C. Related Work

According to, Borrower Misreporting and Loan Performance (Garmaise, 2015), borrowers who reported above-threshold assets faced loan default of around 25 percent with the mean being 20%. Applicants with unverified assets faced delinquency in excess of 40 percent. Other reports find that around 48% of loans exhibited at least one indicator of misrepresentation (Griffin, 2016). Combining these statistics it is clear investors can fall for a good story (Herzenstein, 2011) and misinformed investors can make poor decisions on who to lend to due to poor evaluation on the borrower's creditworthiness (Iyer, 2009).

There is additional damage to the economy when profits are gained by betting against the financial system and potential bad loans. This phenomenon was illustrated in the 2015 film, "The Big Short" (Gardner, 2015) in which men bet against the financial system and made millions of dollars were made on the 2008 housing crisis by betting that against the financial systems that helped provide loans to homebuyers.

D. Outline of Paper

The rest of the paper is organized as follows. Section II is a description of the materials (data) and methods used. Section III contains the results. Section IV outlines potential challenges to completing the work. Section V notes future work. Section VI concludes the paper.

II. MATERIALS AND METHODS

A. Data Collection

The Lending Club dataset was downloaded from https://www.kaggle.com/wendykan/lending-club-loan-data with a file size of 110,111,739 bytes. The data set consists of 887,379 observations of 74 variables. The dataset is described as version 1 and was accessed during the week of 1 May 2016. The files contain complete loan data for all loans issued through the years 2007-2015, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. A data dictionary is provided from the same source that describes variables in the loan data as well as a few that are not present in the data.

B. Data Preparation and Processing

Variables from the superset that would be unknown at the time of loan application submission will be removed. Remaining observations that have N/A or outlandish values will be removed. An example of an outrageous data combination for this data set is when annual income is reported as zero dollars with debt to income ratio being reported as $9999.00. This situation represents where default values have made their way into the data set. These values will skew any evaluation they are apart of. Remaining observations are evaluated for correlation amongst their variables. Variables with categorical only values are removed as correlation is most easily calculated against numeric variables.

R Studio

A digitized brute force approach to check overall correlation between variables can be applied within RStudio quite easily. The data is imported into a data frame and cleaned to the above specifications. The partially processed data presents as 886,877 observations of 8 variables for the numeric correlation. The outcome is shown below.

	loan_amnt	int_rate	installment	annual_inc	dti	revol_bal	revol_util	total_acc
loan_amnt		14.51%	94.50%	33.26%	2.06%	33.36%	11.96%	22.26%
int_rate			13.31%	-7.29%	8.00%	-3.56%	26.91%	-3.85%
installment				32.61%	1.43%	31.66%	13.18%	20.04%
annual_inc					-8.74%	29.56%	3.67%	18.74%
dti						6.73%	8.78%	10.80%
revol_bal							21.62%	18.92%
revol_util								-11.32%
total_acc								

Figure 1. Correlation of 8 numeric variables from all valid observations. Note: Observations with no data or clearly invalid data have been removed.

Eureka

Eureka is software from Nutonian Inc at http://www.nutonian.com/. The modified dataset from R can be imported into Eureqa and Eureqa can run automated evaluations to find potential regression models to explain the data. Evaluations can be based on error ratios, R^2 goodness of fit and/or other methods.

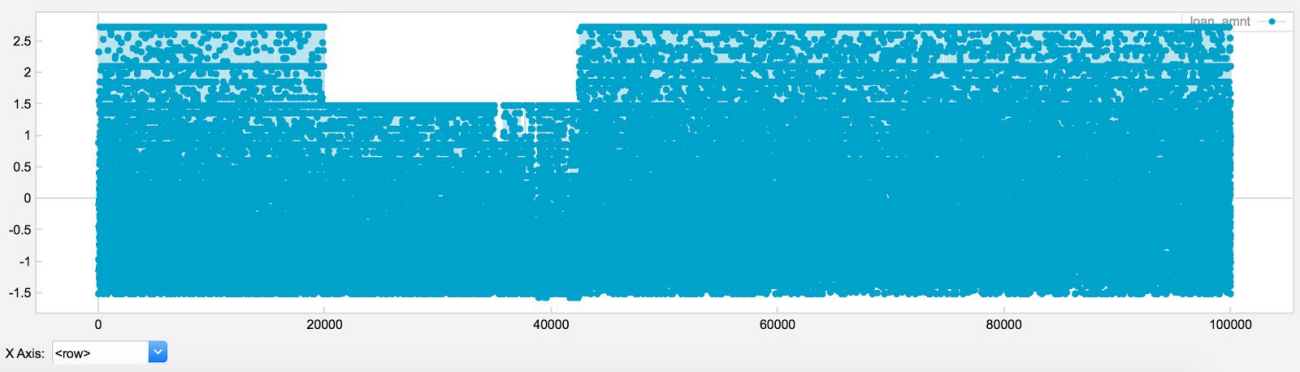

Figure 2. Dot Plot output from Eureqa number of loans vs normalized reported income

The preparation output from Eureka presented in dot plot form shows what appears to be lines drawn through the graph and various levels. This demonstrates that the categorical variables may cause division within the dataset that must be accounted for in the pursuit of a regression model that explains the income.

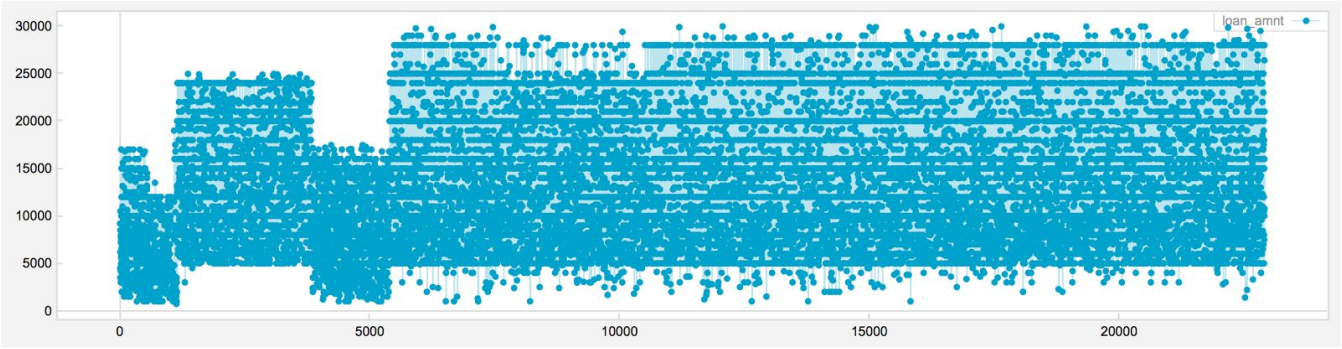

Figure 2. Eureqa data preparation output for loan amount vs all observations in the A1 Subgrade

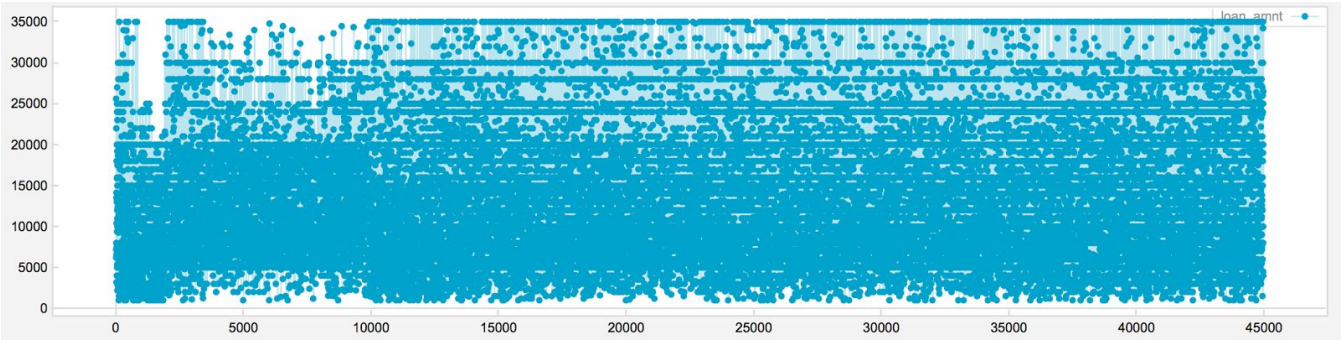

Figure 3. Eureqa data preparation output for loan amount of all observations in the B1 Subgrade

It is readily apparent that even within the categorical breakdown of subgrade the data is further fragmented by other factors. These factors must be identified and their effect on a resulting model analyzed and applied effectively.

III. RESULTS

A. Description of Data Set Created for analysis

Early analysis of the original dataset revealed that not all of the variables would be available or necessary in developing a linear model with a relatively high R^2 value. The remaining variables were partitioned and loaded into Eureqa. Eureqa a proprietary Artificial Intelligence.-powered modeling engine. The software uses evolutionary search to determine mathematical equations that describe sets of data in their simplest form (Nutonian.com, 2016).. Attempts were made to normalize the data based upon the size of the training dataset using with standard normal type conversion for the analysis. The dataset overall remained left untouched for other processes with the option to use the conversion of the data instead of the untreated original. Ultimately the initial broad view of the data required it to be stripped down to 8 numeric variables that could have correlations drawn from their values.

B. Potential Data Science Approach

Our dataset is made up of the 887,379 observations, each observation consists of 74 variables which represents data on a loan application. This data includes all the applicants' information when they apply for a loan without listing personal information that would identify the borrower. After reviewing different data science approaches, the initial approach selected was generating a regression model. This choose provided allowed a predictive element to help identify significant outliers in declared income

suggesting that this income may be fraudulent. Regression is defined as a technique in which a straight line is fitted to a set of data points to measure the effect of a single independent variable. The slope of the line is the measured impact of that variable (Kutner, 1996). The regression model will use several indicator variables to define the outcome variable of annual income. Reasonable values in tested data can be determined by developing a 95% confidence interval around the result and comparing the two to detect outliers. Correlation is the simultaneous change in value of two numerically valued random variables (Kutner,1996). This allows for the testing of various other types of null hypothesis.

C. Preliminary findings

Since unverified loans were less likely to default than verified and sourced verified loans that data was analyzed to determine the number and percent of unverified loans among all defaulted loans in the data set. This was achieved by loading the Lending Club Loan dataset into Tableau and within the Tableau environment using filtered sets to sort the loans that are 2 year delinquent based upon its verification. The unverified loans make up 23% of the defaulted loans while verified and source verified loans made up about 77% (FIGURE[1], Table[1]). This finding was surprising since it stands in contrast with our initial assumption that loan applicants with unverified and potentially fraudulent income would be more likely to default than applicants with verified income.

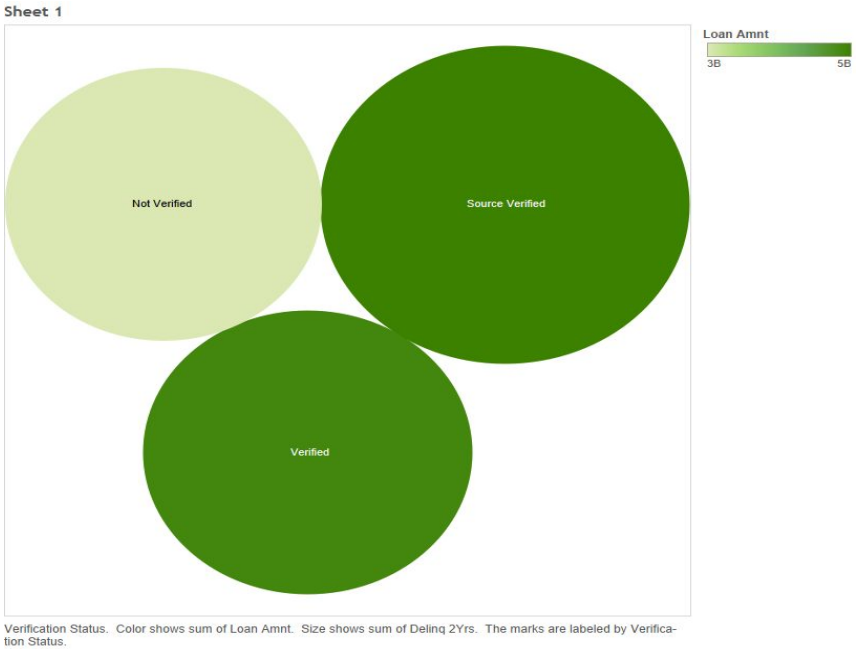

FIGURE4 Default loans as it pertains to Verification Status

Table[1]: Actual Numbers Used to Generate Figure 1

Verification Status	Acc Now Delinq	Delinq 2Yrs	Loan Amnt
Not Verified	897	81,110	2,981,260,350
Source Verified	1,852	110,202	5,107,987,625
Verified	1,680	87,708	5,004,263,975

D. Rationale for using data to address research question

The Lending Club loan data was used in this study for the following reasons: (1) the data set was large with more than 800,000 observations, (2) it was provided by a company that has a significant role in the P2P lending industry, (3) the data set contains a significant number of observations allowing for easy generation of a training set, (4) the data is hosted by Kaggle and for a time will be retrievable by any group seeking to replicate any findings and (5) the data set was available at no cost.

IV. **POTENTIAL CHALLENGES**

The data source reflects a single source of data provided by Lending Club which may not be represented of the entire P2P industry. This presents a challenge where other potentially related variables are unaccounted for in the data set. Another challenge is the correlation of different data sets may not yield a suitable overlap of useful variables. Within the subset of data initially analyzed the 8 numeric variables are clearly grouped by a few of the categorical. Being able to understand the effect that the categorical variables have upon the superset is essentially required if a regression model approach is to be attempted. A simple enough example that demonstrates absent utility is the value of a home in terms of assets. Of larger interest might be demographic data of the consumers. For a more robust correlative model, datasets from different types of loan providers are necessary. While a model can potentially be built using a single source, the opportunity enhances the model and identify other relationships that are present must be pursued.

V. **FUTURE WORK**

Knowing the correlative effects of various variables on one another could be useful in the creation and comparison of models for other lending groups. It may also be possible to find supersets comprised of other institution's datasets combined with this one that can be used in similar analysis which might include regions or other demographic correlative data. The potential to identify fraudulent practices by lenders is also appealing.

Further research should look to dissect this dataset and break it down by each individual sub-grade and other categorical variables to look for the reasoning between the differences in correlation amongst the various subgroups. This could allow several different models to work for each individual sub-grade. One goal is to find the best model for the entire dataset. Statistics using categorical data must also be incorporated as most of the Lending Club dataset is made up of categorical data. Incorporation of other lender's data sets, both traditional and P2P type to further investigate the relationship between the data could have distinct value. Analysis should either confirm or nullify the hypothesis that the differences demonstrated through the lending club data set are equal to self-reported values. Development of the model and process could provide support for multiple echelons. It can give auditors and regulators another tool to apply toward companies and determining if the status quo is at work or if creative data fraud is taking place. It can also be given to the companies themselves to augment their own processes to reduce fraud and overall cost.

Work toward sorting of the data to understand why at certain observations the data is vastly skewed needs to be done to ensure that the data itself hasn't been tampered with in any obvious fashion.

VI. CONCLUSION

Understanding the contradicting nature of the presented data and various papers is vital in knowing

how to approach further analysis of potential fraud. Further study can use the outcome in conjunction with other lenders' data to develop a model to help predict when an unverified applicant's reported income is beyond the threshold of the expected norm. The results of the model can indicate to auditors or governing bodies that something is potentially wrong with the data and more than an automated approach is warranted.

ACKNOWLEDGMENT

Dr Todd Gary PhD. is acknowledged for his insight and instruction in scientific research that improved the final version of our first scientific paper. Sarah Grotelueschen provided insight and candor that improved earlier drafts of this paper. Lending Club is acknowledged for providing the principal data set used throughout this paper and Kaggle.com for hosting the Lending Club data.

REFERENCES

Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent Systems in Accounting, Finance and Management*, *18*(2-3), 59-88.

Adelino, M., Schoar, A., & Severino, F. (2015). *Loan Originations and Defaults in the Mortgage Crisis: The Role of the Middle Class* (No. w20848). National Bureau of Economic Research.

Chaffee, E. C., & Rapp, G. C. (2012). Regulating Online Peer-to-Peer Lending in the Aftermath of Dodd-Frank: In search of an evolving regulatory regime for an evolving industry. *Wash. & Lee L. Rev.*, *69*, 485.

Collins M, (2015, July) The Big Bank Bailout, *Forbes, http://www.forbes.com/sites/mikecollins/2015/07/14/the-big-bank-bailout/#e09c0b33723f*

Duffee, G. R., & Zhou, C. (2001). Credit derivatives in banking: Useful tools for managing risk?. *Journal of Monetary Economics*, *48*(1), 25-54.

Freedman, S., & Jin, G. Z. (2008). Do social networks solve information problems for peer-to-peer lending? evidence from prosper. com.

Gardner, D. (Producer), & McKay, A (Director). (2015). *The Big Short* [Motion Picture]. USA: Paramount Pictures

Garmaise, M. J. (2015). Borrower misreporting and loan performance. *The Journal of Finance*, *70*(1), 449-484.

Green, R. K., & Wachter, S. M. (2005). The American mortgage in historical and international context. *The Journal of Economic Perspectives*, *19*(4), 93-114.Griffin, J. M., & Maturana, G. (2016). Who facilitated misreporting in securitized loans?. *Review of Financial Studies*, *29*(2), 384-419.

Herzenstein, M., Sonenshein, S., & Dholakia, U. M. (2011). Tell me a good story and I may lend you money: the role of narratives in peer-to-peer lending decisions. *Journal of Marketing Research*, *48*(SPL), S138-S149.

Holt, J. (2009). A summary of the primary causes of the housing bubble and the resulting credit crisis: A non-technical paper. *The Journal of Business Inquiry*, *8*(1), 120-129.

Iyer, R., Khwaja, A. I., Luttmer, E. F., & Shue, K. (2009, August). Screening in new credit markets: Can individual lenders infer borrower creditworthiness in peer-to-peer lending?. In *AFA 2011 Denver Meetings Paper*.

Jiang, W., Nelson, A. A., & Vytlacil, E. (2009). Liar's loan? Effects of origination channel and information falsification on mortgage delinquency.*Indiana University-Bloomington: School of Public & Environmental Affairs Research Paper*, (2009-06), 02.

Kaggle.com. Lending Club Loan Data. https://www.kaggle.com/wendykan/lending-club-loan-data , 2016

Karlan, D., & Zinman, J. (2008). Lying about borrowing. *Journal of the European Economic Association*, *6*(2-3), 510-521.

Kearns, D (2016, February, 9th) Pros and Cons of a Cash-Out Refinance, [NerdWallet], https://www.nerdwallet.com/blog/mortgages/refinance-cash-out/

Koller, C. (2010). *Diffusion of innovation and fraud in the subprime mortgage market* (Doctoral dissertation, University of Cincinnati).

Kusisto, L (2015, April) Many Who Lost Homes to Foreclosure in Last Decade Won't Return — NAR, *Wall Street Journal*,
http://www.wsj.com/articles/many-who-lost-homes-to-foreclosure-in-last-decade-wont-return-nar-1429548640

Kutner, M. H. (1996). *Applied linear statistical models* (Vol. 4, p. 318). Chicago: Irwin.

Larrimore, L., Jiang, L., Larrimore, J., Markowitz, D., & Gorski, S. (2011). Peer to peer lending: The relationship between language features, trustworthiness, and persuasion success. *Journal of Applied Communication Research*, *39*(1), 19-37

Light, J (2015, April)Treasury Department: Fannie, Freddie Bailout Wasn't A Loan, *Wall Street Journal,*
http://blogs.wsj.com/developments/2015/04/21/treasury-department-fannie-freddie-bailout-wasnt-a-loan/

Lin, M., Prabhala, N. R., & Viswanathan, S. (2013). Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Management Science*, *59*(1), 17-35.

Mian, A. R., & Sufi, A. (2015). *Fraudulent income overstatement on mortgage applications during the credit expansion of 2002 to 2005* (No. w20947). National Bureau of Economic Research.

Narayanan, A., Shi, E., & Rubinstein, B. I. (2011, July). Link prediction by de-anonymization: How we won the kaggle social network challenge. In *Neural Networks (IJCNN), The 2011 International Joint Conference on* (pp. 1825-1834). IEEE.

Nutonian.com, Eurequa, http://www.nutonian.com/products/eureqa/ , 2016

Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of Default in P2P Lending. *PloS one*, *10*(10), e0139427.

Slattery, P. (2013). Square pegs in a round hole: SEC regulation of online peer-to-peer lending and the CFPB alternative. *Yale J. on Reg.*, *30*, 233.

Yang, Y. (2015). *Analysis and assessment of credit rating model in P2P lending: an instrument to solve information asymmetry between lenders and borrowers* (Doctoral dissertation, Massachusetts Institute of Technology).

APPENDIX

Place figures and tables here. Add figure legends and titles to tables.

FIGURE [1] Default loans as it pertains to Verification Status:
Tableau graphic of variables: Verification Status, Loan Amount (Sum), and Accounts now delinquent (Sum)

Table[1] Default loans as it pertains to Verification Status:
Tableau table of variables: Verification Status, Accounts now delinquent (Sum), Delinquent for two years(Sum), Loan Amount (Sum)

All loan data (sqlite database)
database.sqlite.zip (134.64 MB)
https://www.kaggle.com/wendykan/lending-club-loan-data/downloads/database.sqlite.zip

Data Dictionary (Variable Definitions)
LCDataDictionary.xlsx (20.5 KB)
https://www.kaggle.com/wendykan/lending-club-loan-data/downloads/LCDataDictionary.xlsx

Loan Data (csv)
loan.csv.zip (105.01 MB)
https://www.kaggle.com/wendykan/lending-club-loan-data/downloads/loan.csv.zip