

A Principal Decision: The Case of Lending Club

A THESIS

Presented to the University Honors Program

California State University, Long Beach

In Partial Fulfillment

of the Requirements for the

University Honors Program Certificate

Jonathan Ricardo Guzman

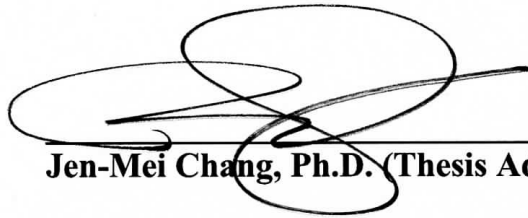
Spring 2015

**I, THE UNDERSIGNED MEMBER OF THE COMMITTEE,
HAVE APPROVED THIS THESIS**

A Principal Decision: The Case of Lending Club

BY

Jonathan Ricardo Guzman

A large, stylized handwritten signature in black ink, featuring several loops and a long horizontal stroke extending to the right.

Jen-Mei Chang, Ph.D. (Thesis Advisor)

Mathematics and Statistics

California State University, Long Beach

Spring 2015

Acknowledgments

I would like to give my deepest thanks to my advisor and mentor Dr. Jen-Mei Chang for her invaluable support in the past year. Her guidance during the research process was insightful and inspiring.

I would also like to thank Nen Huynh for providing me valuable resources and for taking time out of his busy life to help me develop my thesis.

Finally, I want to thank Dr. Tianni Zhou. We could not have continued far into this research without your crucial insight into statistical matters and analyses.

It was a pleasure working with you all.

Table of Contents

Acknowledgments.....	iii
Table of Contents.....	iv
List of Tables.....	v
List of Figures.....	vi
Introduction.....	1
Preliminary Analysis.....	8
Principal Component Analysis.....	14
Linear Regression.....	16
Summary.....	25
Conclusion.....	27
Works Cited.....	29

List of Tables

Table 1: Non-categorical independent variables generated from Lending Club profile features..	10
Table 2: Categorical features and associated dummy variables.....	11
Table 3: ANOVA table for the our multiple regressor model.....	18
Table 4: ANOVA table for the reduced model.....	23

List of Figures

Figure 1: A screenshot of the Lending Club loan browsing page.....	4
Figure 2: A typical profile of an applicant at a glance.....	6
Figure 3: Scatter plot of actualized return rates versus logarithm of loan funded amount.....	12
Figure 4: Scatter plot of return rate versus the total number of accounts under the applicant's name.....	13
Figure 5: Plot of singular values acquired from the singular value decomposition of 119 variable data.....	15
Figure 6: Histogram of residual values resulting from the forward-stepwise selection. Note the prominent right-skew of the values.....	20
Figure 7: Scatter plot of return rates versus residuals. Again, note the extensive cluttering and deviation of points at higher values of residuals.....	20
Figure 8: A quantile-to-quantile (Q-Q) plot of residual data versus standard normal quantiles. Note the deviation from the “perfect” standard normal red-line.....	21
Figure 9: Histogram of residuals generated from the reduced model.....	23
Figure 10: Scatter plot of return rates versus residuals in the reduced model.....	24
Figure 11: Q-Q plot of residuals in order gauge its deviation from a standard normal distribution.....	25

1. Introduction

With the emergence of e-commerce, investors no longer function simply as sources of capital to financial institutions. Traditional customers are now investing through other financial intermediaries or modes. As online lending services continue to grow and develop, investors behave like, and transform into, bank-like entities themselves.

Berger and Gleisner argue that the growth of the Internet has led to a subsequent increase in usage of online financial intermediaries as substitutes to traditional banking systems (Berger and Gleisner 3-6). The social nature of the Internet, in short, has given rise to a more social means of borrowing and lending money.

Hulme and Wright purport this emergence of “social” or “peer-to-peer” lending transforms the investor to an entity who now considers the risks and benefits of potential borrowers crudely and as a whole--without the shroud of a bank, but also without the risk-mediation a bank offers (Hulme and Wright 10). Nonetheless, Hulme and Wright also assert that the borrowers and lenders both enjoy the fact that peer-to-peer lending “...creates the perception that the exchange is experientially real and fundamentally more genuine than experiences in mainstream financial services.”

Coupled with the inherent risk of a more personalized form of financial mediation, Klafft summarizes the experience and result of peer-to-peer engagement: “...[A]n expensive middleman is replaced by a more cost effective online platform...[and] borrowers are given the chance to present their loan case in much detail...that banks with their standardized decision processes

usually do not take in to consideration" (Klaft 1). Overall, Klaft argues that transparency exposes lenders to "significant information asymmetries" which in turn allows peer-to-peer lending platforms to "generate higher returns for investors (compared to traditional bank savings)..." (Klaft 2).

Therein lies the purpose of the following research: People find online, peer-to-peer investment more gratifying than traditional savings investments. Thus, as people continue to make the transition from physical institutions to virtual entities, this paper considers which elements of a particular lending platform make it a lucrative investment. Consider the case of Lending Club--an online peer-to-peer lending platform designed to "create a more efficient, transparent and customer-friendly alternative to the traditional banking system that offers creditworthy borrowers lower interest rates and investors better returns." Prior observations cite the lending platform Prosper as their primary area or subject of research, and whilst arguments concerning the concept of peer-to-peer lending platforms generalize to Lending Club, there exists little research into its interface, user base, and mechanisms. Thus, considering the continuing expansion of peer-to-peer lending, it is worth exploring such facets from the perspective of Lending Club.

Founded in 2007, Lending Club allows a user to issue loans to other users or allows the user to apply for a loan. As a creditor (or lender), the user provides monetary funds upfront to Lending Club; the lender is then allowed to issue portions of this pool of money (called notes) to loan applicants in twenty-five dollar increments. In order to apply for a loan, a user must provide credit history and credit factors to Lending Club itself. Lending Club has the authority on whether or not to list a loan request. Ultimately however, lenders choose the particular applicants to which to issue notes; the lender's choice is contingent on other profile data that loan applicants

provide during that application process. Applicant profile data includes information such as a user's employment information, age, current geographic location of an applicant, and a history of credit information (this is not exhaustive). Once listed, multiple lenders issue notes to applicants until their loans are fully funded within a two-week period (when listings expire); applicants then receive their funding (if fully funded) and payment plans at a thirty-six or sixty month period (options chosen by the loan applicant). Overall, the ability of a lender to glean from a profile an applicant's life and history is the transparency entailed by Klafft's analysis on peer-to-peer lending.

With Lending Club simply acting as a filter of “creditworthiness,” a lender essentially becomes a miniature bank--issuing loans based off of profile factors that will maximize expected returns. Once Lending Club determines that an applicant is “creditworthy,” it issues the applicant an “A” through “G” grade and a 1 through 5 subgrade based off of an applicant's credit history. This grade determines the interest that a borrower pays at the end of a loan period: “A1”-grade loans receive the lowest possible interest rates, whilst “G5”-grade loans receive the highest. Figure 1 shows a list of potential loans and their progress towards fulfillment from the perspective of a lender. According to Lending Club, a grade is calculated by adding a base interest rate (at time of writing, 5.05 percent) and a rate that captures the “risk and volatility” that a lender would face if he or she issued a note to a particular loan applicant. With grade and profile information in mind, a lender makes a determination on which loans to fund.

The aim of this research is to answer the following questions:

1. Which profile variables should we consider as inputs in a model that determines expected returns? Which variables are good predictors of this value?

2. How does Lending Club determine its grading system? What applicant histories does Lending Club use to determine this grade?

Build a Portfolio								
Per Loan: <input type="text" value="\$25"/>								
Filter Loans Save Open								
Exclude Loans already invested in ▼								
<input checked="" type="checkbox"/> Exclude loans invested in								
Loan Term ▼								
<input checked="" type="checkbox"/> 36-month								
<input checked="" type="checkbox"/> 60-month								
Interest Rate ▶								
Keyword ▶								
More Filters ▶								
Update Results								
Minimize All Reset All								
Investment	Rate	Term	FICO®	Amount	Purpose	% Funded	Amount / Time Left	
<input type="checkbox"/> \$0	C 2 12.69%	36	660-664	\$2,000	Other	<div><div></div></div> 98%	\$25 13 days	
<input type="checkbox"/> \$0	A 1 5.93%	36	700-704	\$6,500	Loan Refinancing & Consolidation	<div><div></div></div> 91%	\$550 10 days	
<input type="checkbox"/> \$0	D 5 17.86%	60	660-664	\$10,500	Car financing	<div><div></div></div> 94%	\$575 11 days	
<input type="checkbox"/> \$0	A 5 7.89%	36	665-669	\$20,000	Credit Card Payoff	<div><div></div></div> 57%	\$8,425 8 days	
<input type="checkbox"/> \$0	B 1 8.18%	60	810-814	\$18,000	Loan Refinancing & Consolidation	<div><div></div></div> 92%	\$1,425 10 days	
<input type="checkbox"/> \$0	A 3 6.68%	36	725-729	\$5,000	Loan Refinancing & Consolidation	<div><div></div></div> 69%	\$1,550 11 days	
<input type="checkbox"/> \$0	A 5 7.89%	36	685-689	\$7,000	Loan Refinancing & Consolidation	<div><div></div></div> 60%	\$2,800 10 days	
<input type="checkbox"/> \$0	D 5 17.86%	60	685-689	\$30,000	Small Business	<div><div></div></div> 95%	\$1,250 11 days	
<input type="checkbox"/> \$0	A 4 6.92%	36	765-769	\$35,000	Loan Refinancing & Consolidation	<div><div></div></div> 62%	\$13,200 7 days	
<input type="checkbox"/> \$0	A 1 5.93%	36	740-744	\$10,000	Credit Card Payoff	<div><div></div></div> 63%	\$3,625 10 days	
<input type="checkbox"/> \$0	A 4 6.92%	36	710-714	\$15,000	Loan Refinancing & Consolidation	<div><div></div></div> 48%	\$7,700 9 days	
<input type="checkbox"/> \$0	E 1 18.25%	60	695-699	\$16,000	Small Business	<div><div></div></div> 92%	\$1,225 11 days	
<input type="checkbox"/> \$0	A 5 7.89%	36	665-669	\$15,000	Loan Refinancing & Consolidation	<div><div></div></div> 46%	\$8,100 9 days	
<input type="checkbox"/> \$0	A 3 6.68%	36	675-679	\$8,500	Loan Refinancing & Consolidation	<div><div></div></div> 58%	\$3,500 10 days	
<input type="checkbox"/> \$0	A 5 7.89%	36	700-704	\$8,500	Loan Refinancing & Consolidation	<div><div></div></div> 50%	\$4,200 10 days	

Figure 1: A screenshot of the Lending Club loan browsing page.

The first question arises as a consequence of lending money through peer-to-peer platforms (through Lending Club in particular): Put simply, “What is that 'noise' in the data?” Lending Club provides a filtering system to quickly expedite the loan process; lenders can filter in or out loan listings that meet (or do not meet) certain qualifications on the user side. Filter options not only include profile data, but grade is also a potential filter.

Lending Club provides public access to sets of data and tables concerning loan statistics. One such table details actualized returns versus the grade of particular loan applicants across

completed loans. According to Lending Club, data indicates that grades “C” through “E” historically yield higher eventual returns. Thus, this research also determines whether a lender should simply consider the grade of a loan applicant or a combination of profile attributes aside from the grade.

The second question arises from attempts to answer the first: If grade is the only factor a lender should consider, then what determines grade? The research in this paper operates under the assumption that the “risk and volatility” of a loan applicant is calculated using information on his or her profile. While Lending Club is not explicit about how it calculates this facet of the grade, credit history certainly factors into this rate, and some credit history is actively portrayed on a user's loan profile or in the statistics gathered by Lending Club. Therefore, in determining which profile factors are indicative of potential returns, if these credit factors arise, then the determination is that grade serves as the “best” indicator of expected returns. Note that this analysis also incorporates the possibility that grades and profile combinations together form indicators.

This paper first tackles the history of Lending Club loans--fulfilled or otherwise. Lending Club allows users access to three spreadsheets-worth of data that incorporates every loan ever listed on the Lending Club website. This paper will first detail and explain the descriptive statistics of select profile factors across a multitude of loans. These profile factors are selected based off of what a lender would typically consider when issuing loans--factors like the length of employment of an applicant, the debt-to-income ratio of an applicant, and utilization of existing bankcards. The research shows the relation between these factors versus the actualized return rates of borrowers--not only as a whole group, but also amongst categories of applicants determined through selected profile options in the loan application process (for example,

applicants by state or applicants by purpose-of-loan); patterns and tendencies of data versus actualized returns are indicative of possible significance. Overall, factors that exhibit trends lend themselves more readily to the mathematical process called principal component analysis.

The second section of this paper will detail how and why these select factors are the “best” indicators of expected returns or otherwise. By converting spreadsheet data into arrays of numbers and normalizing said arrays, the research determines the principal components of profiles encapsulated in Lending Club's historical data--the factors significant to determining potential returns of lenders as an output of profile inputs. Principal component analysis itself only determines the n -number of significant factors. Figure 2 shows an expanded loan profile with potential factors listed. Note that the a borrower’s grade is the primary feature listed.

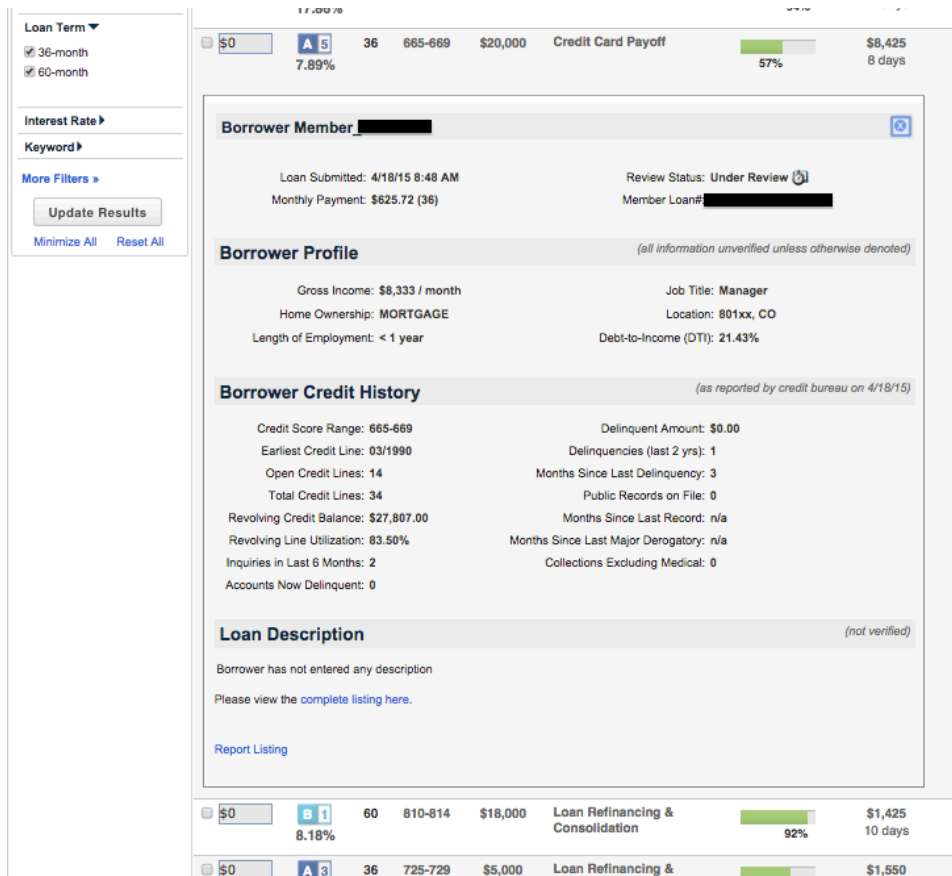


Figure 2: A typical profile of an applicant at a glance.

This paper will finally delve into the process of linear regression: Taking combinations of profile factors n at a time, linear regression determines statistical individual correlative and joint correlative values to the factor combinations, and it also assigns weights to these factors which gauge the effect (increase or decrease) and intensity (by how much) each component has on expected return.

2. Preliminary Analysis

2.1 Loan data

The profile of a loan applicant contains 100 features that a loaner considers prior to issuing a note. It is essential to consider a smaller set of these variables in order to facilitate a viable conclusion.

Firstly, we consider only “completed” loans--loans that had reached 36- or 60-months of activity. Furthermore, we consider loans under “policy code” 1--loans that are publicly available on the lending platform. We ignore profile features that elicit no substantial bearing on the expected return of a particular loan. For example, a user-provided description as to the usage of a requested loan provides information that is already captured in the “purpose” parameter provided by Lending Club; the purpose feature is kept over the description since loan purpose is simple to quantify. Overall, we consider features that are inherently continuous (such as monetary values) and features that are easily quantifiable.

On the whole, Lending Club loan data entails features with missing data. For the most part, this lack of data is not caused by an applicant's negligence (i.e., his or her inability to answer a question during the application process). The bulk of the missing data results from Lending Club's review of an applicant's credit report. The reason behind this loss is unknown. Yet while not all applicant profiles are missing credit information, the issue is systemic enough to warrant the removal of these features from the component and regressive analyses. It is also important to note that most of these columns of data would not be included on an applicant's

profile anyway. After eliminating these data, we filter out any remaining profiles that have missing entries of data. This filtering process, as opposed to filtering out all profiles with missing data without eliminating features, maximizes our sample size and strengthens the viability of our eventual model.

Lastly, when appropriate, categorical variables (such as the aforementioned “purpose” feature or “grade” feature) are converted to dummy variables when category options are preset. Assuming there are n -number of options to choose from in a particular category feature, our analyses convert this information into $n - 1$ binary variables, where 1 means the applicant has selected a particular option and 0 if otherwise. Table 1 gives a summary of the non-categorical variables considered, and Table 2 gives a summary of the categorical variables and associated dummy variables.

In all, these adjustments are necessary to make sense of the data in the context of Lending Club. These changes reflect how a typical loaner chooses an application to fund--only considering a few key features from the 100 available. Following the aforementioned measures, we consider a sample of 16985 loans from the original pool of 17723 complete loans. Of these loan profiles, we consider thirteen continuous-value features and four category features (geographic region, purpose, home ownership status, and grade) with appropriate dummy variables incorporated. These features reflect what a typical lender considers before issuing a note to a potential borrower; they also contain minimal amounts of missing information that could otherwise adversely affect our eventual model.

Profile Feature	Variable	Description
log(funded_amnt)	X_1	The total amount funded to the loan applicant, converted into logarithmic values.
int_rate	X_2	Applicant's interest rate, determined by Lending Club.
log(annual_inc)	X_3	Applicant's annual income, converted into logarithmic value.
delinq_2yrs	X_4	The number of thirty-day past-due incidences of delinquency in the applicant's credit file in the past two years.
dti	X_5	Applicant's debt-to-income ratio.
emp_length	X_6	The number of years the applicant was employed at time of applying.
high_fico	X_7	The upper boundary of range the applicant's FICO belongs to.
open_acc	X_8	The applicant's number of open credit lines.
pub_rec	X_9	The applicant's number of derogatory public records.
pub_rec_bank	X_{10}	The applicant's number of public record bankruptcies.
revol_bal	X_{11}	The applicant's total revolving credit balance.
revol_util	X_{12}	The applicant's total usage of revolving credit.
total_acc	X_{13}	The applicant's total number of credit lines currently on the borrower's file.

Table 1: Non-categorical independent variables generated from Lending Club profile features.

2.2 Descriptive statistics

The descriptive statistics concerning the pertinent profile features justify the use of the aforementioned principal component and multiple linear regression analyses. Primarily, we consider the interaction of these features with the expected return rate of a loan--the ratio between the amount paid back over the amount loaned.

Looking at scatter plots of various pertinent features versus expected return rate, it is clear that no discernable pattern emerges. Preferably, a scatter plot would show a negative (downward-sloping) or positive (upward-sloping) correlation between the independent variable

and return rates. Figures 3 and 4 demonstrate this lack in correlation actualized returns and X_1 and X_8 , respectively. The data points themselves appear to only occur along specific lines. This is partly due to the nature of these profile features: Most of these features quantify discretely or behave like discrete values; values like “funded amount,” which entails the continuous value of a person's requested loan, behave discretely when their logarithm is calculated.

Profile Feature	Variables	Description
addr_state	NE; NW; W	The applicant’s geographic region at time of applying--options are Northeast, Northwest, West, and Midwest.
home_ownership	MORT; OWN; RENT	Applicant’s home ownership status at time of applying.
purpose	HOME_IMPROV; CREDIT_CARD; DEBT_CONSOL	The applicant’s purpose for borrowing.
grade	A; B; C; D; E; F	Applicant’s profile grade, as calculated by Lending Club.

Table 2: Categorical features and associated dummy variables.

Looking at scatter plots of various pertinent features versus expected return rate, it is clear that no discernable pattern emerges. Preferably, a scatter plot would show a negative (downward-sloping) or positive (upward-sloping) correlation between the independent variable and return rates. Figures 3 and 4 demonstrate this lack in correlation actualized returns and X_1 and X_8 , respectively. The data points themselves appear to only occur along specific lines. This is partly due to the nature of these profile features: Most of these features quantify discretely or behave like discrete values; values like “funded amount,” which entails the continuous value of a person's requested loan, behave discretely when their logarithm is calculated.

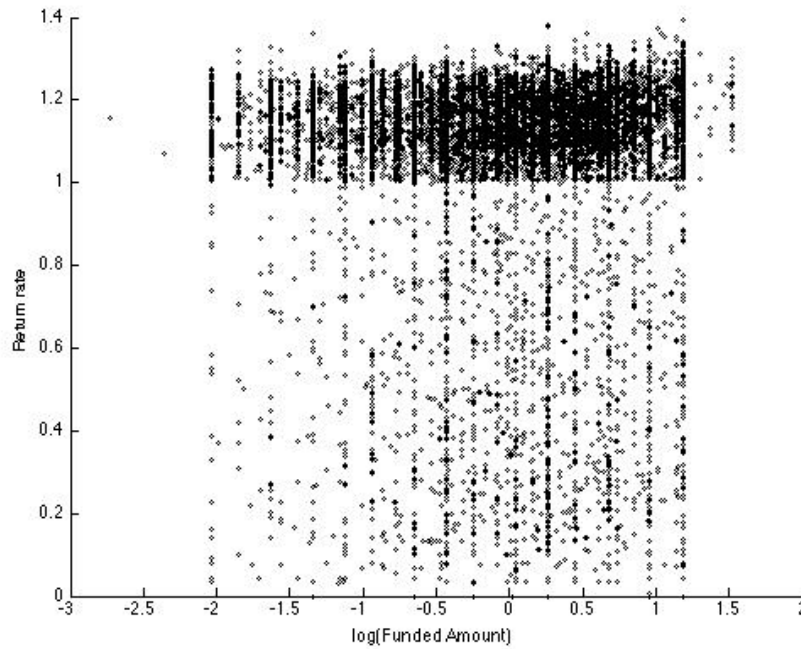


Figure 3: Scatter plot of actualized return rates versus logarithm of loan funded amount.

The underlying idea in these plots, however, is that a regression in one variable is not an adequate model. One can discern the amount unexplained variation between this hypothetical “best-fit” line and actualized return rates. This is further supported by the sheer amount of features on the an applicant's profile: Return rate must be the response variable in a multi-dimensional system. This further implies that a regression in multiple variables can help explain the variation present in a simple regression.

In either case of multiple linear regression, where we consider solitary linear independent variables or quadratic interaction terms, the task of obtaining the “best” model simply by iterating through all possible combinations of variables is arduous and costly in the sense of time. In the former case, twenty-eight variables taken in combinations up to twenty-eight at a time yields 2^{28} or 268435456 models. In the latter case, the thirteen non-categorical variables yield

ninety-one quadratic interaction terms--on top of the existing twenty-eight variables, yielding 119 terms; this amounts to 2^{119} or approximately 6.46×10^{35} models!

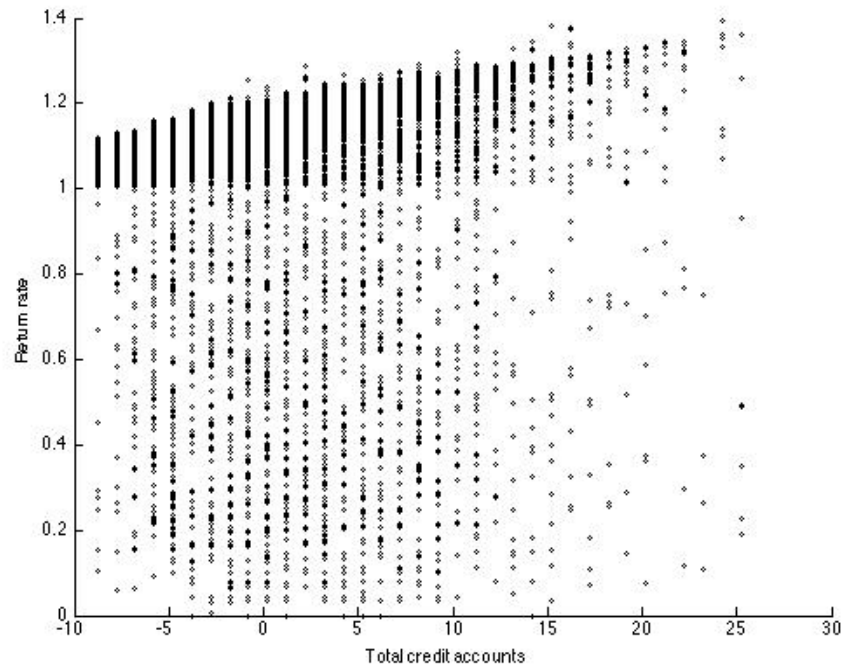


Figure 4: Scatter plot of return rate versus the total number of accounts under the applicant's name.

3 Principal Component Analysis

Before employing selection criteria to generate the “best” subset of regressors, we first consider the “best” number of mathematical bases to represent the aforementioned data.

Each instance of a loan can be thought of as a vector in 119-dimensional space. Physically, it is impossible to visualize these points in this state. However, if one were to project these vectors down into a space spanned by orthonormal bases (preferably a span which encompasses vectors in two- or three-space), one could ascertain characteristics of the data based on how the projected data points cluster together.

In order to uncover these principal bases, we find and uncover hyperplanes, in iterated steps, such that at each step we minimize the square of the residuals--the square of the distance between a loan data point and this hyperplane; we then project this data down to this determined hyperplane. Algebraically, this amounts to calculating the singular value decomposition of the profile data and projecting the data down to an r number of bases that maximize the explained variability (minimizes the residuals) of the profile features.

Figure 5 shows the plot of singular values. This graph indicates that the number of orthogonal bases should be somewhere around fifteen or sixteen--the “elbow” of the graph. By means of MATLAB, the singular value decomposition shows that the number of bases needed to explain ninety-five percent of the data (about two standard deviations of a standard normally distributed data) is four; in order to explain 99.7 percent of the data (about three standard deviations), fourteen bases are needed.

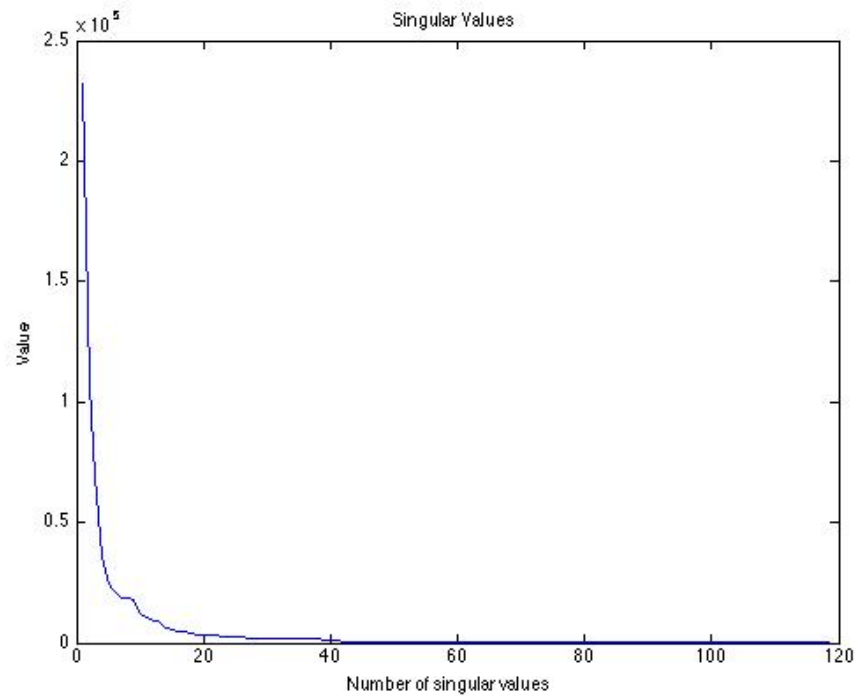


Figure 5: Plot of singular values acquired from the singular value decomposition of 119 variable data.

The purpose of this preliminary analysis is to better guide our eventual model selection. Preferably, our selection of a regression model will yield around fourteen variable and parameter estimations; this model will thus have a high explanatory power--well-fit against actualized loan data points. Otherwise, the results of the PCA will function as model selection criteria.

4 Linear Regression

4.1 Forward-stepwise selection

We consider iterative and selection algorithms to narrow down the number of choices and to better utilize computational resources. Using the MATLAB programming language, and corroboration through the statistical computing language R, the following model is chosen using a forward-stepwise selection method:

1. Start with no variables in the model. Begin by selecting from the k number of variables and fit simple linear regression models to these variables individually. The variable that yields the highest F-statistic (our selection criterion, though others can be used) is the candidate for entry into the model; if this statistic is higher than a pre-determined critical score, the variable enters the model.

With this new variable in the model, the algorithm now repeats the following steps until no more variables can be added or removed from the model:

2. Fit a multiple linear model with the existing model-variables and a new variable one at a time. Again, the variable that yields the highest F-statistic is the candidate for entry and once again be higher than a preset critical score to enter.

3. One at a time, remove variables (excluding the variable that was immediately added before) from the model. Obtain the appropriate F-statistics and determine the lowest; this determines the candidate for deletion. If the F-statistic falls below a pre-determined value, the variable is dropped.

4.2 Multiple linear regression with second-order interaction terms

A linear regression with second-order interaction terms captures more variation in the system as variables can sometimes influence each other. In the context of Lending Club, this is especially true given how some variables like X_1 (*open_acc*) and X_2 (*total_acc*) are closely related.

Using an “entry significance level” of .10 and “exit significance level” of .15, the forward-stepwise process yields the following “best” model under the given conditions:

$$\begin{aligned} RETURN_RATE = & 1.0660 + 0.0093X_2 + 0.0344X_3 - 0.0004X_7 - 0.0051X_{11} - 0.0052X_{13} - 0.0202W + 0.0393A + \\ & 0.0341B + 0.0324C + 0.0265D + 0.0195E + 0.0139F - 0.0155X_3^2 - 0.0004X_6^2 + 0.0000X_7^2 + 0.0004X_{11}^2 - \\ & 0.0318X_1X_{10} - 0.0026X_1X_{11} + 0.0218X_2X_{10} + 0.0297X_3X_4 + 0.0013X_3X_8 + 0.0004X_3X_{12} + 0.0012X_4X_5 + \\ & 0.0590X_4X_9 + 0.0021X_7X_{10} + 0.0000X_7X_{11} - 0.0000X_7X_{12} - 0.0067X_8X_{10} - 0.0002X_8X_{11} - 0.0001X_8X_{12} + \\ & 0.0011X_{10}X_{12} - 0.0157X_{10}X_{13} + ADJUSTMENT \end{aligned}$$

Since there are interaction terms that contain variables with no linear representation, *ADJUSTMENT* represents the manual entry of these linear terms:

$$ADJUSTMENT = 0.0026X_1 + 0.0031X_4 + 0.0004X_5 + 0.0001X_6 + 0.0002X_8 - 0.0096X_9 + 0.0159X_{10} - 0.0001X_{12}$$

Jointly, the results of the search implementation and accompanying adjustments appear to be significant, within 90%-confidence level--as detailed by the analysis of variance (ANOVA) in Table 3. The adjusted R^2 statistic--the ratio of variation explained by the model and total variance--is approximately 0.0242. This value is crucial in analyzing the overall effectiveness and explanatory power of the model.

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F-stat	Crit-F ($\alpha = 0.10$)
Model	40	22.2694	0.0556735	11.5124	1.296
Error	16944	819.4052	0.04836		
Total	16984	841.6747			

Table 3: ANOVA table for the our multiple regressor model.

Interestingly, only a single geographic location affects the expected return rate of a borrower: An application originating from western states would seemingly decrease the expected return by .0202 percent, all else held constant. The presence of all grades runs counter to our initial assumptions. Prior to this analyses, we presumed that grades “C” through “E” would yield a higher return rate compared to any other grades. This comes from overall grade trends and loan maturity data provided by Lending Club. Lastly, there are contradictory effects for certain interaction terms as well--namely, the single highest contributor to return rate X_4X_9 (two-year delinquencies multiplied by the number of public derogatory records) with marginal effect of 0.0590 percent; according to this model, the higher the number of delinquencies on a borrower's record, the higher his or her return rate will be. In fact, the overall effect of X_4 is theoretically $0.0031 + 0.0297X_3 + 0.0012X_5 + 0.0590X_9$ --which is positive given any value of these variables. This contradiction persists with the effect of the square of the logarithm of annual income--negative, despite the positive effect of the associated linear term.

Linear regression entails certain assumptions that the data may or may not follow. In order to ascertain the state of the profile data, we consider the residuals, or errors in estimation, and their distribution relative to the standard normal distribution. One central tenet of linear regression holds that residuals must be distributed normally; otherwise, the aforementioned model may not be an appropriate least-squares estimation of expected returns.

4.3 Analysis

As mentioned before, the forward-stepwise selection process ensures that the chosen variables in the model are jointly significant as per an appropriate F-test. However, other measures indicate that this model is far from “good.”

The ANOVA values indicate that the model has minimal explanatory power in the context of Lending Club. With a sum of errors of approximately 819.4052, where total sum of squares is 841.6747, the amount of explained variation lies at about 2.42 percent--far less than half, which most consider an appropriate benchmark of model viability.

Beyond this, the proceeding figures are also troubling. Figure 6 clearly shows that the residuals do not elicit a normal distribution; the data is too skewed to the right. This skew towards higher residual values is further highlighted in Figure 7, where the variation in value increases as residuals increase. Though, the Q-Q plot illustrated in Figure 8 between residuals and normal quantiles is the final indication of this heavy skew. The graph is heavily off-center; this is a result of the inherent skew in residuals, but overall illustrates a failure in the primary linear regression assumptions.

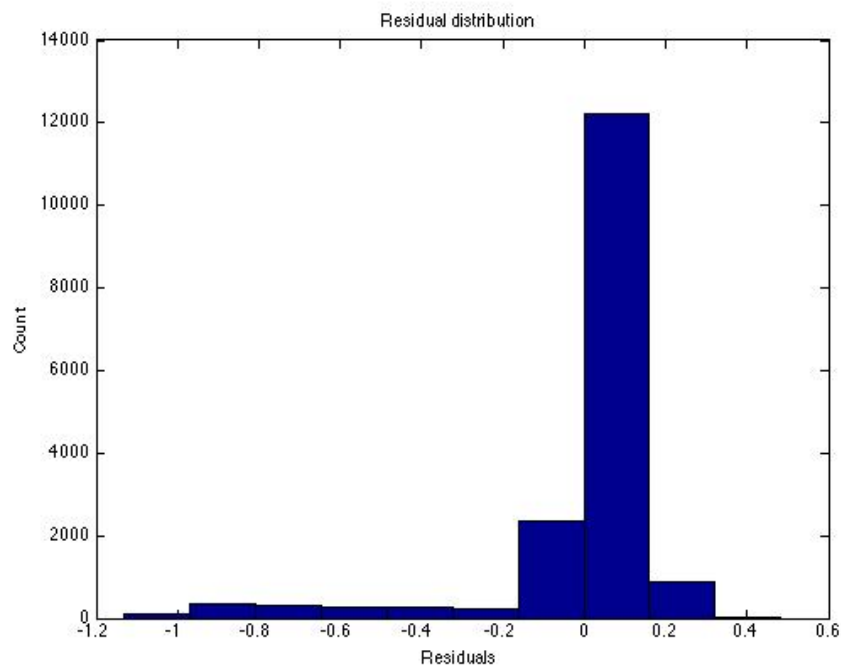


Figure 6: Histogram of residual values resulting from the forward-stepwise selection. Note the prominent right-skew of the values.

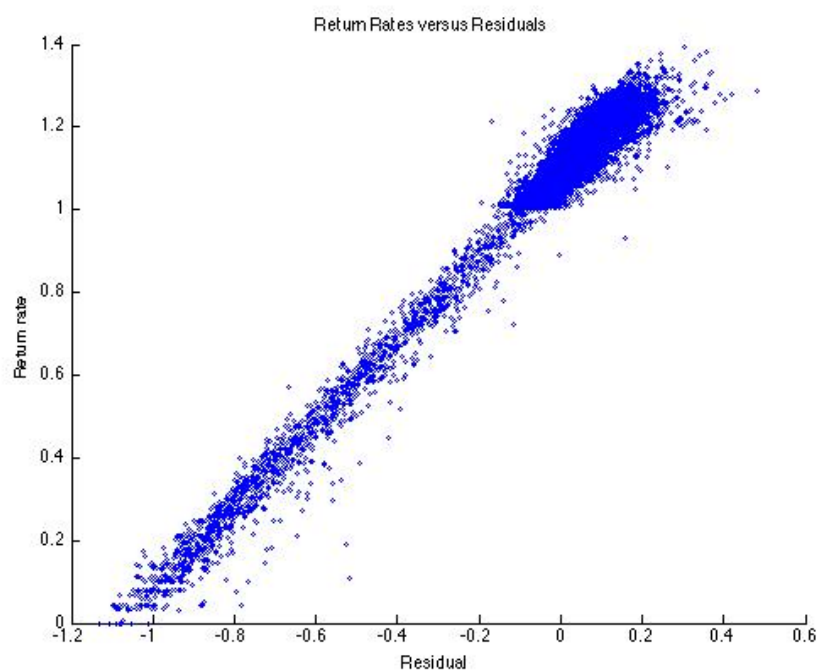


Figure 7: Scatter plot of return rates versus residuals. Again, note the extensive clustering and deviation of points at higher values of residuals.

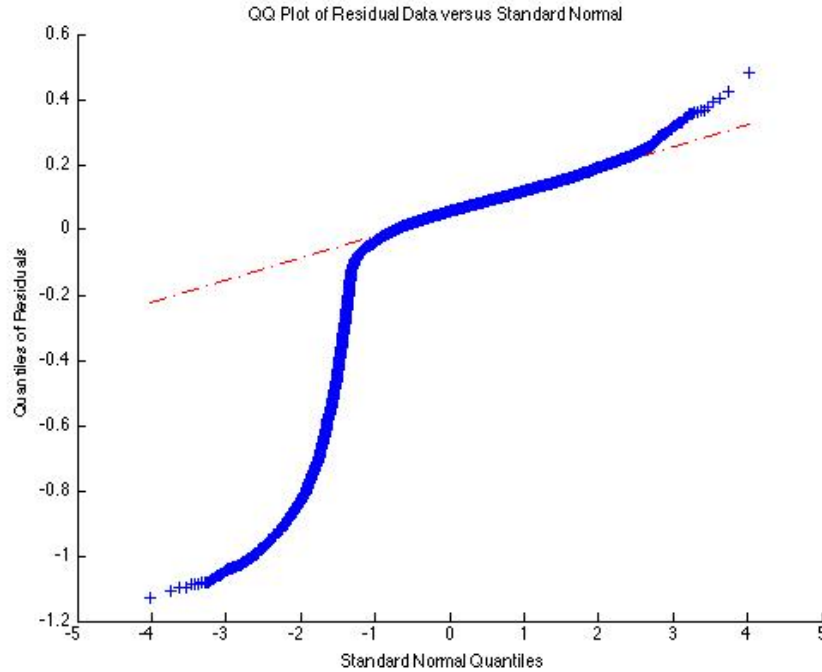


Figure 8: A quantile-to-quantile (Q-Q) plot of residual data versus standard normal quantiles. Note the deviation from the “perfect” standard normal red-line.

4.4 Model re-selection

The forward-stepwise selection procedure yields a model that does not adequately represent lending profile data. Thus, we consider the appropriate number of regressors acquired previously through the PCA procedure. From the forty variables in the model mentioned above, we consider fourteen of these variables to formulate a better model--one that hopefully captures 99.7 percent of the variance in the data.

However, this process creates a logistical problem in terms of managing computer resources. The total amount of combinations generated by taking forty elements fourteen at a time is 23206929840. To alleviate the computational stress of calculating all these number of regression models, we randomly generate a set of models and apply selection criteria to identify the one which has the highest explanatory power (adjusted R^2 value).

Iterating through these random selections, it is clear that adjusted R^2 does not rise to the level of our preliminary model. A glance at these reduced models seems to indicate that the maximum of this value lingers below two percent. We consider this randomly selected model:

$$\begin{aligned} RETURN_RATE = & 1.0921 + 0.00898X_2 + 0.02701X_3 - 0.00059342X_6 - 0.0048918X_9 - 0.003382X_{11} - 0.0043634X_{13} \\ & - 0.019436W + 0.0045078D - 0.013712X_3^2 - 0.037153X_1X_{10} - 0.0024133X_1X_{11} + 0.0018103X_7X_{10} - \\ & 0.0000087X_7X_{12} - 0.0070256X_8X_{10} + ADJUSTMENT \end{aligned}$$

Again, certain linear terms are not coupled with their second-order interaction terms. As such, *ADJUSTMENT* represents these addition as compensation:

$$ADJUSTMENT = 0.0010229X_1 - 0.00039416X_7 - 0.0010713X_8 - 0.0000945X_{12}$$

Within a 90%-confidence interval, the analysis of variance in Table 4 shows that this re-selected model has less explanatory power than the first--at 1.77 percent. The most striking features of this model are the reductions in the effects of some variables. Most of the associated variable coefficients indicate positive or negative effects of less than a percent change in expected returns--barring variables in the adjustments. "D" is now the only profile grade a lender considers under this model. This falls in line with our initial evaluation of grade, but this model also excludes "C" and "E." There are contradictory effects (higher FICO score implies lower return rate), but the magnitude and presence of these effects are lessened in the reduced model.

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F-stat	Crit-F ($\alpha = 0.10$)
Model	18	15.7647	0.875816667	17.987	1.296
Error	16966	825.91	0.048680302		
Total	16984	841.6747			

Table 4: ANOVA table for the reduced model.

The residuals generated by ordinary least squares estimation of the reduced loan data gives the same insight as before: Figure 9 details the distribution of residuals generated by the reduced model; residual values appear highly dependent on the value of expected returns. Figure 10 plots these residuals versus actualized return rates; higher expected returns is correlated with higher residual values. In fact, this is the same relation present with the residuals in the preliminary model. Furthermore, the same right-skew is present in both preliminary and reduced models. Finally, the Q-Q plot of residuals shown in Figure 11 in the reduced model indicates a violation in the primary least-squares assumptions: The residuals do not exhibit a normal distribution.

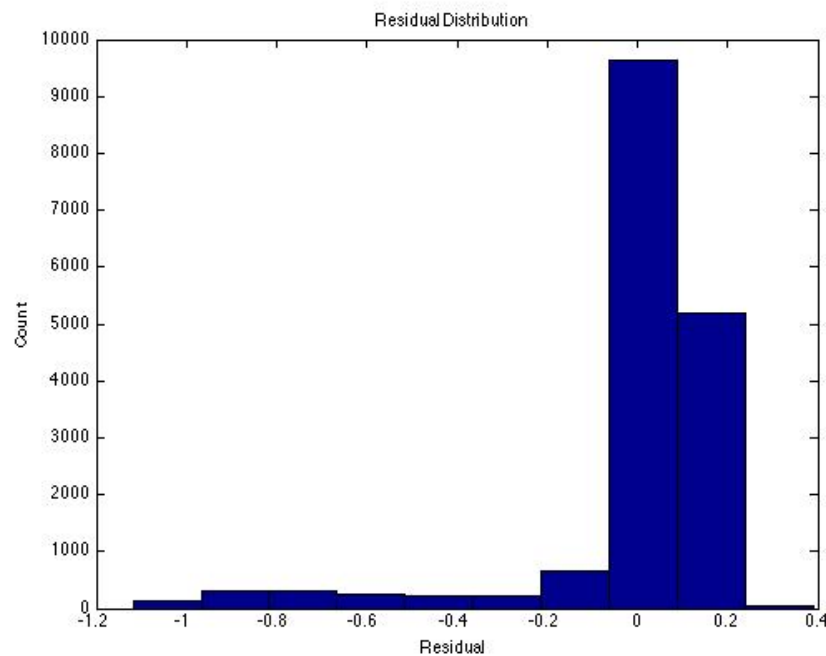


Figure 9: Histogram of residuals generated from the reduced model.

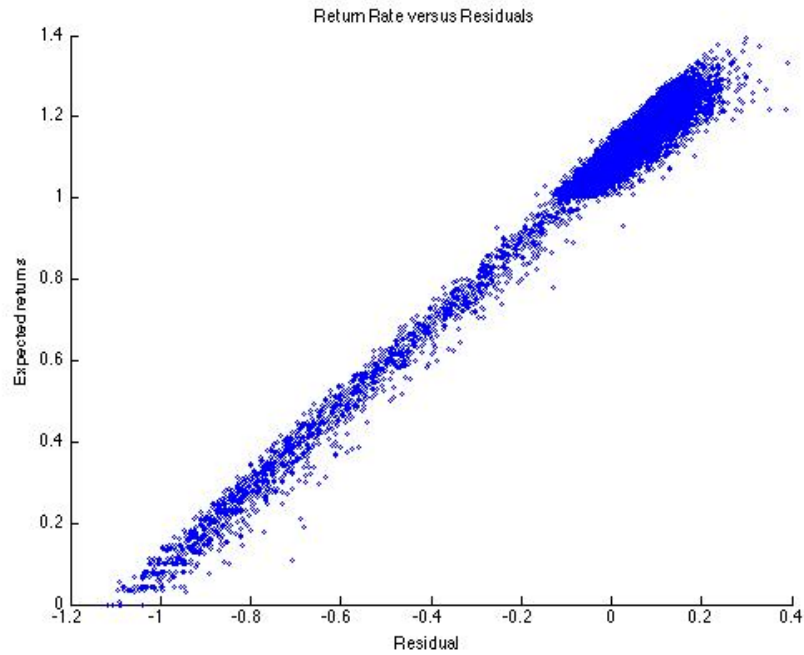


Figure 10: Scatter plot of return rates versus residuals in the reduced model.

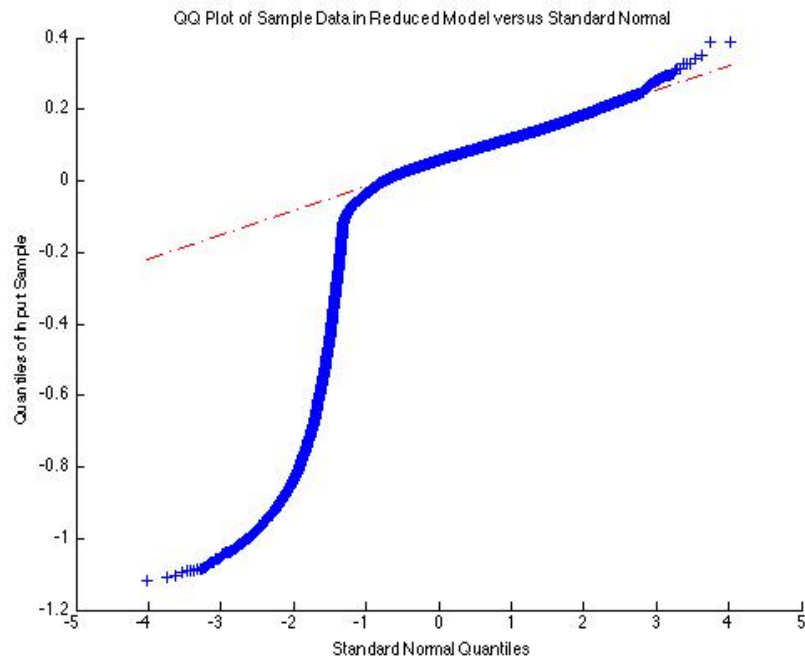


Figure 11: Q-Q plot of residuals in order gauge its deviation from a standard normal distribution.

5 Summary

From the 100 features present on a Lending Club loan applicant profile, our research narrows this number down to twenty-three features--with emphasis on viability and data availability. As some of these features are binary inputs, we convert these features to an appropriate number of dummy variable features; this process yields 119 viable features to consider in a multiple linear regression model with second-order interactions.

An initial principal component analysis indicates that four features explain ninety-five percent of the variability in the data and fourteen features explain 99.7 percent. While the process does not indicate which of these features are the principal components of the data, the results of this initial handling of data serves as additional selection criterion in the regression analysis.

Our preliminary model names forty variables of the viable 119 via a multiple linear regression. While this model meets our expectations with regards to grade and interest rate, it fails to explain about ninety-seven percent of the variability in the loan profile data. Furthermore, an analysis of residuals indicates that the data does not meet the primary assumptions of ordinary least-squares estimation--namely, normally distributed residuals.

Finally, in order to consolidate the PCA and multiple linear regression model, we randomly remove twenty-six from our model in order to find a fourteen-variable model which captures 99.7 percent of the data. The resulting reduced model once again failed to accommodate

all but 1.77 percent of the loan data variability. The model further illustrates the aforementioned failure in OLS estimation assumptions.

6 Conclusions

One aim of this research was to identify which profile features a lender should consider when making a loan. The forward-stepwise yielded forty variables out of a possible 119 variables and their interactions. Again, there effects on expected returns are jointly significant, yet the model in which they belong fails to accurately capture the data's variance. It is also possible that the forward-stepwise algorithm did not yield the single “best” model out of 2^{119} number of models; the design of the algorithm ignores potential models that branch out from addition or subtraction of variables even if these models would eventually have the highest F-statistic compared to the algorithm's output.

This research partially answers the second question posed in the beginning of this report: Lending Club seems to determine interest rates (and in turn, grades) by relying on the FICO score aspect of the borrower. In fact, the correlation coefficient between interest rate and high-end FICO score is -0.75243: A negative correlation is appropriate since a higher FICO score indicates high trustworthiness and ability-to-pay on the borrower's end--thus a lower interest rate. The only drawback to this answer is that this research does not take into account all profile variables; it only considers the ones determined by our preliminary choosing and the choices determined by our selection processes.

Overall, the answers to these two questions are hindered by the inherent nature of the profile data. Our research assumes the data behaves linearly, but the mechanisms behind return rates appear to be of a higher order than one. However, along the vein of linear regression, there

is room for improvement. Preferably, with enough resources, we could forego the selection process by iterating through all 2^{119} potential regression models and selecting from the total list; this would give a definitive assessment of the first question albeit in a monumental amount of time. Furthermore, we can extend our analysis to the entirety of the loan data: All original 100 variables and all completely loans with complete data.

In lieu of such a comprehensive analysis, the grade profile feature (in reality, the interest rate feature), as advertised on the Lending Club website, continues to function as the foremost indicator of expected returns.

Works Cited

- Berger, Sven C., and Fabian Gleisner. "Emergence of Financial Intermediaries in Electronic Markets: The Case of Online P2P Lending." *Verband der Hochschullehrer fUr Betriebswirtschaft e.V.* 2.1 (2009): 1-27. Print.
- Hulme, Michael K., and Collete Wright. "Internet Based Social Lending: Past, Present and Future." *Social Futures Observatory*. (2006): 1-115. Web. 4 April 2014.
- Klaft, Michael. "Online Peer-to-Peer Lending: A Lenders' Perspective." *International Conference on E-Learning, E-Business, Enterprise Information Systems, and E-Government*. 2008. Las Vegas: IEEE, 2008. Web. 4 April 2014