Modeling the interest rate charged by the Lending Club

Alan Arnholt

Apr 10, 2016

Contents

1 Introduction

The Lending Club ("The Leader in Peer to Peer Lending: Loans and Investing Lending Club" 2016) says it uses technology and innovation to reduce the cost of traditional banking and to offer borrowers better rates and investors better returns. Literature on the Lending Club web site states that the interest rate the Lending Club charges borrowers is based on a club base rate with an adjustment for risk and volatility with further modifiers based on the amount of the loan and the length of the loan.

Modeling the relationship between interest rate and other recorded data allows the reader to gain an understanding behind the so called "proprietary model" used by the Lending Club to set interest rates for the loans it administers. The analysis and model used in this paper suggest that the interest rate charged by the Lending Club is indeed related to the amount and the length of the loan. Individuals with identical FICO scores can use the model in this paper to predict the interest rate the Lending Club would charge them based on a combination of the applicant's monthly income, open credit lines, and inquiries in the last six months.

2 Methods

2.1 Data Collection

Data used in this paper was originally downloaded from

https://spark-public.s3.amazonaws.com/dataanalysis/loansData.csv

February 16, 2013, and again April 10, 2016, using the `R` programming language (R Core Team 2016). It is not clear from the available information how, from whom, or when the data was collected, nor is it clear

what entity or organization did the collecting. Thirty loans were removed that had either questionable values or missing data. Loans were removed when their recorded data conformed to the "Decline Criteria" given at the bottom of the https://www.lendingclub.com/public/how-we-set-interest-rates.action web page.

2.2 Exploratory Analysis

Exploratory analysis was performed by examining contingency tables, density plots, and scatter-plots of the "cleaned" data. The quality of the "cleaned" data was also evaluated for additional discrepancies, and none were noted. To correct the positive skew of monthly income, a base 10 logarithm was applied to monthly income. Added-variable (partial-regression) plots as described in Fox, Weisberg, and Fox (2011) were used in the selection of appropriate variables. Diagnostic plots were used to assess different models including Box-Cox transformations on the response variable (interest rate) as described in Kutner et al. (2005).

2.3 Statistical Modeling

Standard multivariate regression techniques such as those described in Fox, Weisberg, and Fox (2011) and Kutner et al. (2005) were used to develop a model to predict the interest rate of loans awarded by the Lending Club.

2.4 Reproducibility

All analyses performed in this paper can be reproduced by running the original `.Rmd` file with RStudio, assuming the link to the original data remains current and the contents thereof remain unchanged. The `R` packages `car` (Fox and Weisberg 2016), `ggplot2` (Wickham and Chang 2016), `knitr` (Xie 2016b), `rmarkdown` (Allaire et al. 2016), and `bookdown` (Xie 2016a) will need to be installed on the user's computer. Since `bookdown` is being actively developed and is not yet on CRAN, you will need to install `bookdown` from GitHub by typing the following at the `R` prompt:

```
devtools::install_github("rstudio/bookdown")
```

3 Results

The data used to develop the final model includes information on interest rate (`IR`), amount requested in dollars (`AR`), monthly income in dollars (`MI`), number of open credit lines (`OCL`), number of inquiries in the last six months (`IL6M`), loan length in months (`LL`), and a measure of the creditworthiness of the applicant (`FICO`). There were no missing values in the "cleaned" data, which had 2470 loans. Since the distribution of monthly income was skewed right, a log base 10 transformation was applied to monthly income. Variables were added based on partial regression plots and residual analyses. The linear relationship between the square root of the interest rate and the amount of money requested can be seen in Figure 1.

Although the final model includes variables that may measure similar quantities (confounding), the highest variance inflation factor was 12.31 for the variable `OCL`. All other variance inflation factors were less than 10, suggesting multicollinearity is not a significant problem with the final model (Fox, Weisberg, and Fox (2011) and Kutner et al. (2005)). The coefficients in the final model also make sense and are in agreement (sign wise $\pm$) with how the Lending Club claims to award its loans.

The final model used was

$$\begin{aligned}\sqrt{\text{IR}} = \beta_0 + \beta_1 \text{AR} + \beta_2 \text{log10(MI)} + \beta_3 \text{OCL} + \beta_4 \text{OCL}^2 + \beta_5 \text{IL6M} \\ + \beta_6 \text{IL6M}^2 + \beta_7 \text{f(LL)} + \beta_8 \text{f(FICO)} + \beta_9 \text{f(AR:LL)} + \varepsilon \end{aligned} \quad (1)$$

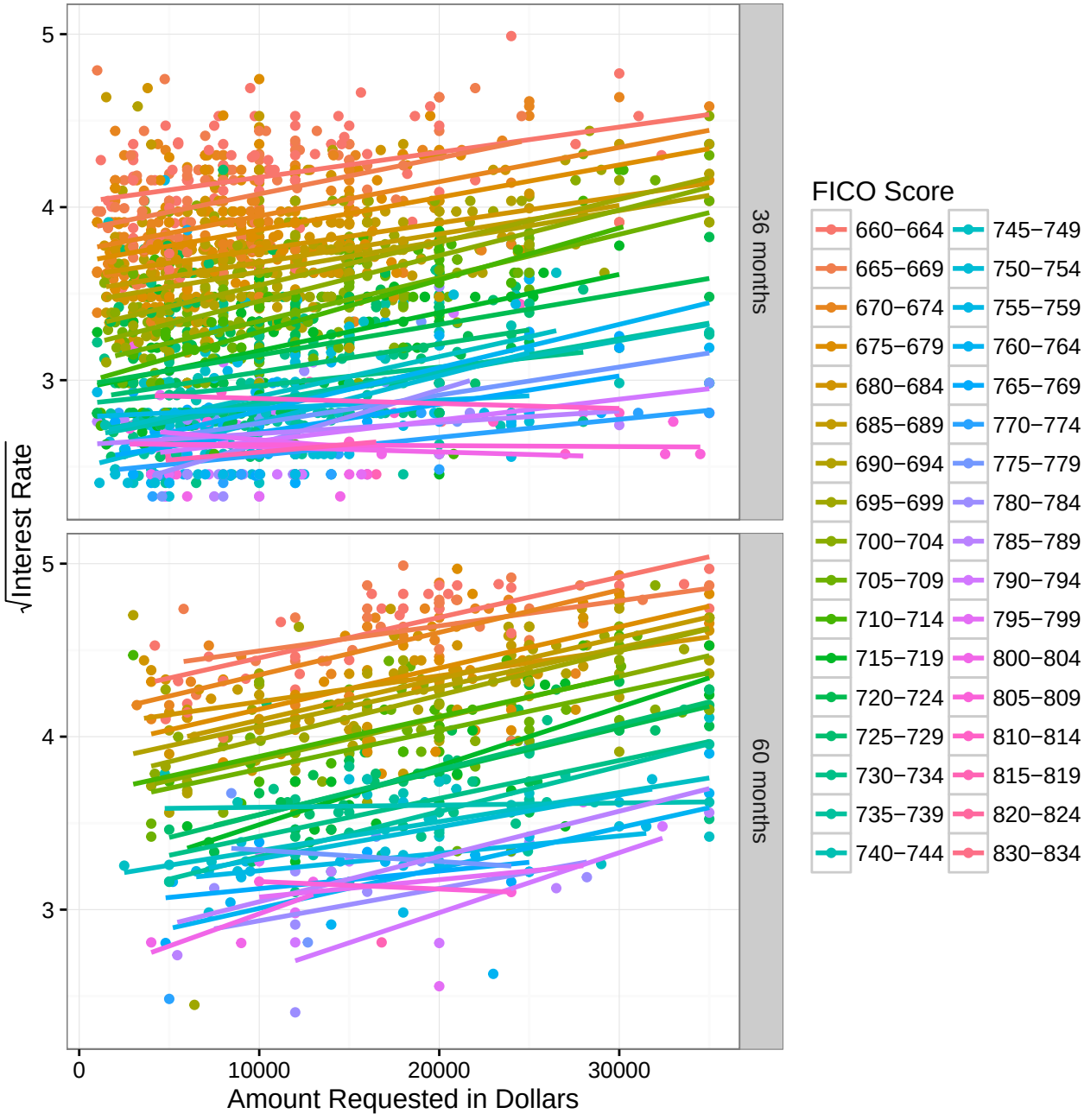

Figure 1: The top panel shows lines obtained from regressing the square root of the interest rate on the amount requested for 36 month loans and the bottom panel shows lines obtained from regressing the square root of the interest rate on the amount requested for 60 month loans. The points and lines are color coded according to FICO scores. The FICO score legend is shown on the right side of the Figure. Higher interest rates generally correspond to lower FICO scores, and the interest rates increase with the dollar amount requested. Interest rates are generally higher for all levels of FICO scores for 60 month loans versus 36 month loans.

Table 1: ANOVA table for the full model fit using ordinary least squares

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Amount.Requested	1	88.6195	88.6195	1491.3542	0.0000
log10(Monthly.Income)	1	16.8119	16.8119	282.9226	0.0000
Open.CREDIT.Lines	1	1.9664	1.9664	33.0919	0.0000
I(Open.CREDIT.Lines^2)	1	5.8905	5.8905	99.1299	0.0000
Inquiries.in.the.Last.6.Months	1	27.9475	27.9475	470.3211	0.0000
I(Inquiries.in.the.Last.6.Months^2)	1	6.9382	6.9382	116.7606	0.0000
Loan.Length	1	67.4247	67.4247	1134.6725	0.0000
FICO.Range	33	492.6554	14.9290	251.2355	0.0000
Amount.Requested:Loan.Length	1	0.2785	0.2785	4.6872	0.0305
Residuals	2428	144.2770	0.0594	NA	NA

The variables `f(LL)`, `f(FICO)`, and `f(AR:LL)` are factors for loan length (2 levels 36 months and 60 months), credit score (34 levels), and the interaction between amount requested and the loan length, respectively. The error term ε is assumed to follow a normal distribution with mean 0 and constant variance. A graph of the residuals versus the fitted model, shown in Figure 2, shows a constant variance for the majority of the range of the fitted values, suggesting the fitted model satisfies the assumptions required for inferential techniques to work with ordinary least squares.

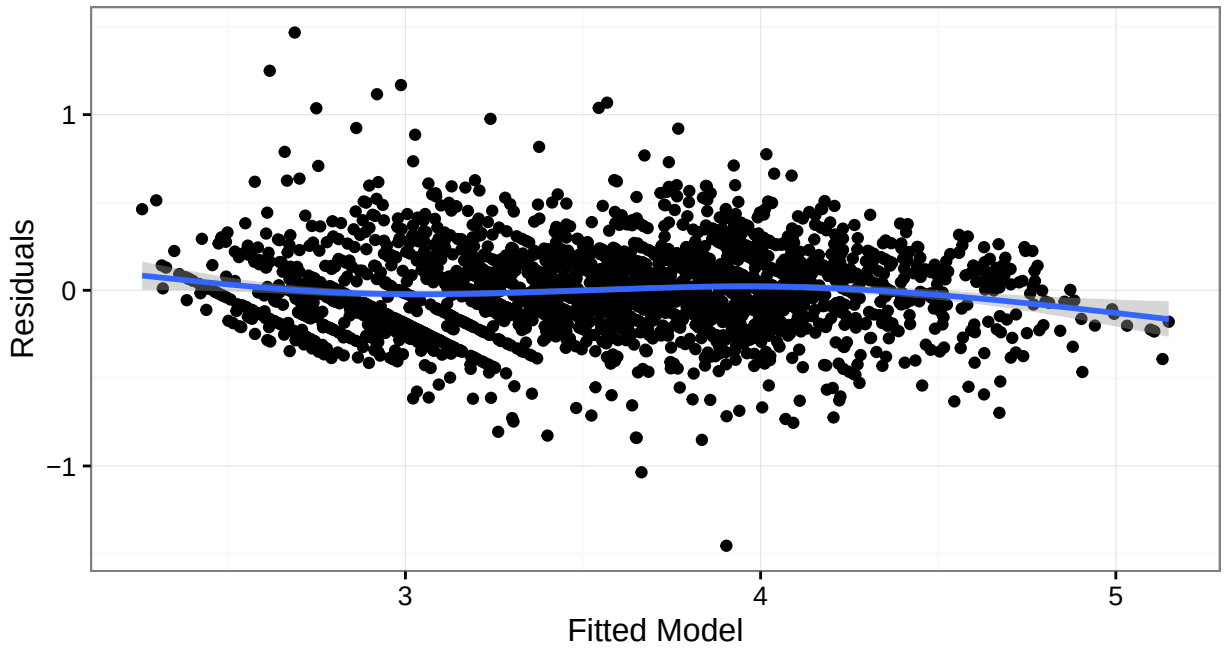

Figure 2: Residuals versus fitted model

There is a highly statistically significant relationship (p-value < 0.0001) between the square root of interest rate and all of the variables in model (1) with the exception of the interaction between the amount requested and the loan length which has a p-value of 0.0305. See Table 1 for complete ANOVA results.

4 Conclusions

Since the goal of the analysis is to predict interest rates, a table showing the mean interest rate, $E(\text{IR}_h)$, given a vector of inputs h is given along with lower (`LB`) and upper (`UB`) confidence bounds for 95% confidence intervals for each $E(\text{IR}_h)$ in Table 2. There is a clear relationship between amount requested, length of loan, and the interest rate charged as evidenced by Figure 1. The expected mean interest rate, $E(\text{IR}_h)$, for a 36 month loan where the values of the input vector h are all at the 0.50 quantile of their respective distributions is 7% (the second row of Table 2). The expected mean interest rate for the same values of the input vector h for a 60 month loan is 9.27% (the fifth row of Table 2). Similar comparisons can be made by studying the values in Table 2 for changes in FICO scores, monthly incomes, amount requested, open credit lines, loan length, and the number of inquiries in the last six months. The reader should note that the confidence intervals reported in Table 2 are individual (not family wise) 95% confidence intervals for the expected mean interest rate computed from an appropriate back transformation so that values are reported on the same scale as the original measurements instead of the square root of the interest rate.

The model used to develop Table 2 has an R^2_{adj} value of 0.828. However, base interest rates change with market conditions and the model in this paper may not work as well for loans made in time periods other than when the data in this paper was obtained.

5 References

Allaire, JJ, Joe Cheng, Yihui Xie, Jonathan McPherson, Winston Chang, Jeff Allen, Hadley Wickham, Aron Atkins, and Rob Hyndman. 2016. *Rmarkdown: Dynamic Documents for R.* http://rmarkdown.rstudio.com.

Fox, John, and Sanford Weisberg. 2016. *Car: Companion to Applied Regression.* https://CRAN.R-project.org/package=car.

Fox, John, Sanford Weisberg, and John Fox. 2011. *An R Companion to Applied Regression.* 2nd ed. Thousand Oaks, Calif: SAGE Publications.

Kutner, Michael H, Chris Nachtsheim, John Neter, and William Li. 2005. *Applied Linear Statistical Models.* Boston: McGraw-Hill Irwin.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

"The Leader in Peer to Peer Lending: Loans and Investing Lending Club." 2016. Accessed April 8. https://www.lendingclub.com/.

Wickham, Hadley, and Winston Chang. 2016. *Ggplot2: An Implementation of the Grammar of Graphics.* https://CRAN.R-project.org/package=ggplot2.

Xie, Yihui. 2016a. *Bookdown: Authoring Books with R Markdown.* https://github.com/rstudio/bookdown.

———. 2016b. *Knitr: A General-Purpose Package for Dynamic Report Generation in R.* http://yihui.name/knitr/.

Table 2: Each row of the first six columns represent different vectors of values (h) passed to the fitted model. The predicted mean interest rate %, $E(\text{IR}_h)$, for each vector h is shown in the column labeled $E(\text{IR})$. The lower and upper bounds on a 95% confidence interval for $E(\text{IR}_h)$ are labeled LB and UB, respectively. The values selected for all quantitative variables are the 0.25, 0.50, and 0.75 quantiles of their respective distributions.

FICO	MI	AR	OCL	LL	IL6M	E(IR)	LB	UB
755-759	3500	10000	9	36 months	0	7.10	6.72	7.49
755-759	5000	10000	9	36 months	0	7.00	6.62	7.38
755-759	6800	10000	9	36 months	0	6.91	6.54	7.30
755-759	3500	10000	9	60 months	0	9.38	8.92	9.86
755-759	5000	10000	9	60 months	0	9.27	8.81	9.74
755-759	6800	10000	9	60 months	0	9.17	8.71	9.65
685-689	5000	10000	9	36 months	0	12.77	12.46	13.08
755-759	5000	10000	9	36 months	0	7.00	6.62	7.38
785-789	5000	10000	9	36 months	0	6.53	5.98	7.11
685-689	5000	10000	9	60 months	0	15.78	15.37	16.20
755-759	5000	10000	9	60 months	0	9.27	8.81	9.74
785-789	5000	10000	9	60 months	0	8.73	8.07	9.42
755-759	5000	10000	7	36 months	0	7.28	6.90	7.67
755-759	5000	10000	9	36 months	0	7.00	6.62	7.38
755-759	5000	10000	13	36 months	0	6.74	6.37	7.13
755-759	5000	10000	7	60 months	0	9.59	9.12	10.07
755-759	5000	10000	9	60 months	0	9.27	8.81	9.74
755-759	5000	10000	13	60 months	0	8.97	8.51	9.45
755-759	5000	6000	9	36 months	0	6.54	6.17	6.92
755-759	5000	10000	9	36 months	0	7.00	6.62	7.38
755-759	5000	17000	9	36 months	0	7.84	7.44	8.25
755-759	5000	6000	9	60 months	0	8.66	8.20	9.14
755-759	5000	10000	9	60 months	0	9.27	8.81	9.74
755-759	5000	17000	9	60 months	0	10.38	9.91	10.86
755-759	5000	10000	9	36 months	0	7.00	6.62	7.38
755-759	5000	10000	9	36 months	0	7.00	6.62	7.38
755-759	5000	10000	9	36 months	1	7.60	7.21	8.00
755-759	5000	10000	9	60 months	0	9.27	8.81	9.74
755-759	5000	10000	9	60 months	0	9.27	8.81	9.74
755-759	5000	10000	9	60 months	1	9.96	9.49	10.45
755-759	5000	10000	9	36 months	0	7.00	6.62	7.38
755-759	5000	10000	9	60 months	0	9.27	8.81	9.74