Data Analysis Assignment 1: Coursera 2013.

By: Dr. Steven J. Rigatti MD

This analysis was completed as part of the Data Analysis course on coursera.com in February of 2013. The purpose of this particular assignment is to identify correlations in the Lending Club dataset. This dataset contains information regarding loans made on the peer-to-peer lending site Lending Club. The interest rates on the loans facilitated through this site are determined by the applicants' characteristics such as employment history, debt to income ratio, FICO score, and loan term. Specifically, the purpose is to determine if there are characteristics that are significantly correlated with the interest rate after taking the FICO score into account.

Methods:

The dataset was downloaded at this site:_. Analyses were carried out in R version 2.15 on the Windows 7 operating system. Standard linear regression[1] and ANOVA[2] techniques were utilized in the construction of the analysis.

The data in its initial form had several shortcomings that needed to be redressed. First, the outcome variable, interest rate, was encoded as a factor with 275 levels. With this may levels between 5.42 and 24.89%, it made more sense to treat this as a continuous variable. The same held true for the debt to income ratio, which was originally encoded as a factor with 1669 levels, and FICO score with was originally encoded as a factor with 38 levels between 640 and 800. The FICO score variable was an index of FICO score ranges (640-644, 645-649, and so on). For the purposes of analysis the FICO scores were transformed into a numeric variable corresponding to the lowest score in the indicated range. The transformation was accomplished with this command in R:

```
loans$FICO<-trunc(as.numeric(gsub("-",".",loans$FICO.Range)))
```

The transformed variables were also preserved in their original factor format for the purposes of utilizing ANOVA techniques.

Analysis:

The majority of the loans made (1952) had a term of 36 months, while 60-month loans numbered only 548. An overlapping histogram (figure 1) of the interest rate segregated by loan term showed that, as would be expected, the interest rate distribution for 36 month loans (mean: 12.12) was lower than for 60 month loans (16.41). A two-sample t-test was used to demonstrate that the mean interest rate for the 2 groups was significantly different (t-score: -22, p<0.001). Similar t-tests were carried out on other variables to determine if there were differences in these characteristics between the 36 and 60-month loan term groups. Significant differences were found for loan amount, monthly income and, as stated, interest rate. Notably, the mean FICO score was not significantly different between the 2 groups (p=0.53, 95% CI: -4.3 to 2.2).

The next step in the analysis was to look for correlations between the interest rate and other variables in the dataset. This was done using the “lm” function in R. The relative strength of these correlations was adjudicated based on the adjusted R-squared measure. This information was used later to inform the construction of multiple linear regression models. Because of the distinct differences in the data between the 36 and 60 month loan term groups, each of the correlations was run 3 times, once for the 36 month group, once for the 60 month group and once for the total sample. Table 2 presents the relevant findings.

Regression Models

Two different regression models were constructed, one to predict interest rate for the whole dataset using loan term as a variable, the other to predict the interest rate for only 36 month loans. Variables were chosen based on the adjusted R-squared measure as determined in the initial analysis. Each time a variable as added to the models, the summary statistics were checked again to see if the new variable added incremental value, based on adjusted R-squared. The results are presented in Table 3.

To summarize, for the dataset as a whole, the FICO score and loan term accounted for a large amount of the variability in the interest rates of the loans being offered. This is illustrated in Figure 2. As additional variables were added to the model, diminishing returns were quickly apparent. However, the amount of the loan requested and the debt-to-income ratio variables did retain some residual utility. In the data subset limited to 36-month loans, the FICO score retained a very high level of correlation with interest rate, though the loan amount did still retain some additional value.

Discussion

This data analysis demonstrated the high correlation of FICO score and loan term with offered interest rates in the Lending Club peer-to-peer web site. These variables were so strongly predictive that it was difficult to improve further upon models based solely on those criteria. Likely this is due to the fact that the FICO score is already a robust measure of creditworthiness, i.e. the ability to pay back borrowed money and therefore already captures much of the information a lending party may be interested to know about a borrower.

Table 1: Mean (sd) of key variables by loan term

	Total (n=2500)	36-month term (n=1952)	60-month term (n=548)	p-value*
Loan Amount ($)	12000(7345)	10334(679)	17942(8030)	<0.001
Monthly Income ($)	5689(3963)	5533(4119)	6242(3294)	<0.001
Revolving Credit Balance ($)	15244(18308)	14706(18980)	17185(15550)	0.002
Debt to Income Ratio	15.38(7.5)	15.28(7.56)	15.74(7.32)	0.19

FICO score	706(35)	706(35.3)	707(34.1)	0.533
Interest Rate (%)	13.07(4.18)	12.13(3.68)	16.41(4.13)	<0.001

*p-value for difference between means for 36 vs. 60-month terms by two-sample t-test

Table 2: Adjusted R-squared (p-value*) for correlation with Interest Rate

	Total (n=2500)	36-month term (n=1952)	60-month term (n=548)
Loan Amount ($)	0.11(<0.001)	0.018(<0.001)	0.11(<0.001)
Monthly Income ($)	0(NS)	0.0003(0.009)	0.014(.0026)
Revolving Credit Balance ($)	0.003(.002)	0(NS)	0.014(.003)
Debt to Income Ratio	0.030(<0.001)	0.032(<0.001)	0.029(<0.001)
FICO score	0.50(<0.001)	0.61(<0.001)	0.66(<0.001)
Open Credit Lines	0.007(<0.001)	0.002(0.02)	0.026(<0.001)
Home Ownership	0.001(0.006)	0.001(0.012)	0.001(0.022)
State	0(NS)	Omitted	Omitted
Loan Length (36 vs. 60 mo)	0.18(<0.001)	NA	NA

Table 3: Multiple regression models

All Data:	Variables considered	Adjusted R-squared
Model 1	FICO score	0.5
Model 2	FICO score + loan term	0.69
Model 3	Model 2 + Debt-to-income ratio	0.69
Model 4	Model 3 + loan amount	0.74
Model 5	Model 4 + monthly income	0.75
36 month loans:		
Model 1	FICO score	0.61
Model 2	FICO score + loan amount	0.66
Model 3	Model 2 + Debt-to-income ratio	0.66
Model 4	Model 3 + open credit lines	0.67

Figure Caption:

Figure 1: Histogram of interest rates. The 30 and 60 month loan term interest rates are plotted overlapping.

Figure 2: Scatter plot of interest rates vs. FICO score for 30 and 60 month loans, with a regression line for each in the corresponding color.