Final project for 140.754
ADVANCED METHODS IV 140.754

Abstract

The Lending Club (`https://www.lendingclub.com/`) has a data set of 2,500 peer-to-peer loans with interest rates records and other information regarding these loans. The goal of this project is to determine variables that are important at inferring the interest rate of a loan beyond the FICO score. In addition, to ensure maximum reproducibility this project has been compiled in an R package *lcollado754*.

1 Introduction

The 2,500 peer-to-peer loans from the Lending Club were analyzed in this project. As can be seen in Figure **??** the FICO score is associated with the interest rate. Overall, the higher the FICO score, the lower the interest rate assigned. The challenge is to then determine if other variables in this data set can help explain differences in interest rates for persons with the same FICO score.

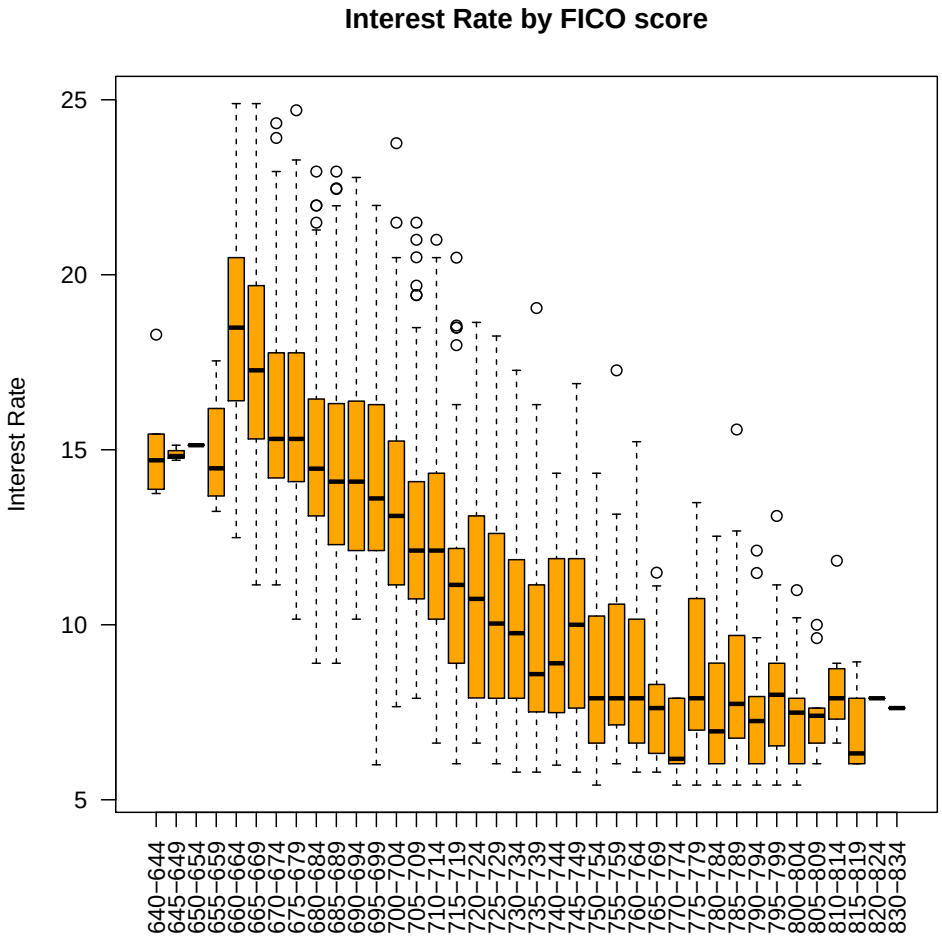

Figure 1: Boxplots of the interest rate by FICO score category. Higher FICO scores lead to lower interest rates.

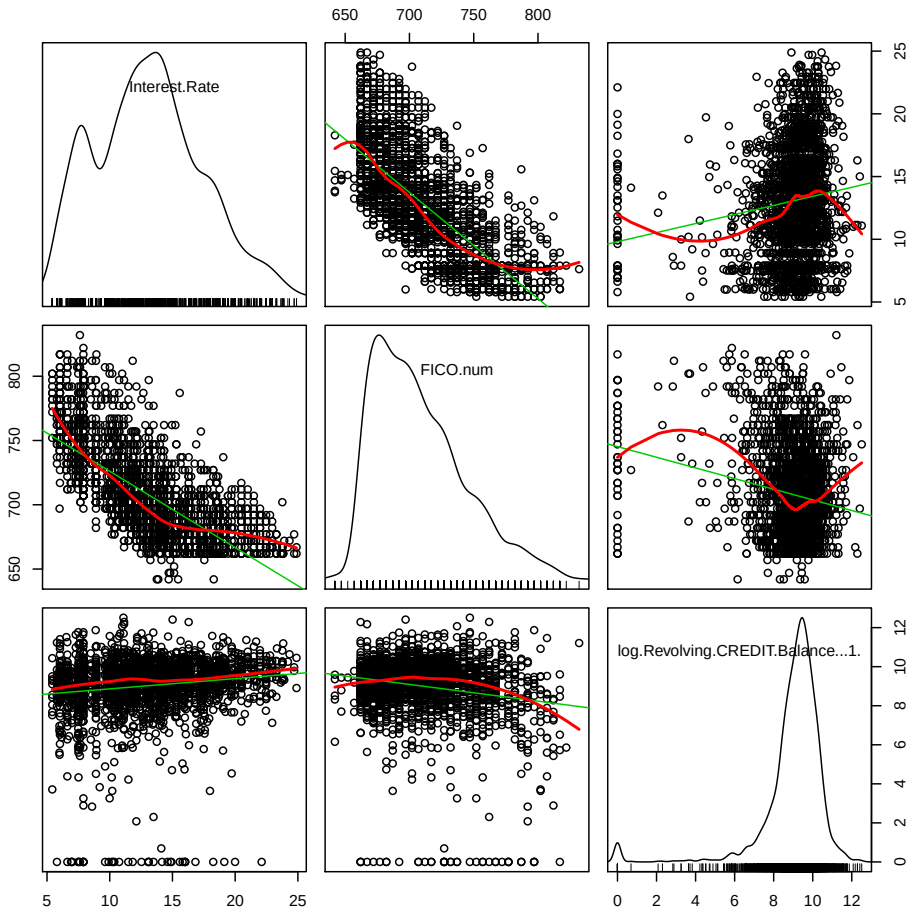

Figure 2: Scatterplot matrix between the interest rate, the FICO score and the logarithm of the the revolving credit balance. The FICO score is shown as numbers taken from the mean value of the categories. The logarithm of the revolving credit balance was taken after adding 1 for ease of transformation purposes.

To achieve the goal of determining which variables explain the interest rate once the FICO score has been accounted for we performed an extensive exploratory data analysis. Some variables, like whether the state is in the East or West of the United States had no relation whatsoever.

From this analysis we also determined that it would be best to reduce the complexity from some of categorical variables. For example *loan purpose* was reduced to *credit card, debt consolidation, or other*. This was because the other categories were very similar in terms of their relation to interest rates and FICO scores.

The log revolving credit balance seems to have a more complicated relationship than the other variables with both the FICO score and the interest rate as shown in Figure **??**. It is notably better to have a low revolving credit balance as long as it's not 0. The distribution of the log revolving credit balance is highly compated around 9. However, it is important to note that at the end range, higher values lead to higher FICO scores and lower interest rates.

2 Methods

The distribution of the interest rate is rather bell-shaped as shown in Figure **??** and so were other numerical variables. It thus made sense to use linear regression methods despite the fact that the domain of the interest rate is bounded in $[0, 100]$. Two general methods were used: multiple linear regression and random forests. Random forests have been shown before to outperform linear models in terms of prediction accuracy and they can be used for regression-type problems and thus seem like a good logical method to compare the linear model to.

In order to compare the two methods, the data set was split into 70% for training purposes and 30% for evaluation purposes. Models were fit on the training data and cross validated within it. Then the robust mean squared prediction error on the evaluating data set was used to determine which model performed best.

2.1 Linear Model

A linear model was fit with all the variables (after pre and post processing). This model was compared against the minimal model using only the FICO score to predict the interest rate and a the model resulting from step-wise model selection (forward and backward) using AIC. The models were then compared to determine which of the three performed the best.

To further evaluate the resulting model, a 10 fold cross validation using 5 replicates was performed. In addition, regular model diagnostic plots were created to evaluate that the linear model assumptions were being met.

2.2 Random Forest

First a random forest with 800 trees was fit with all the variables using the training data. The resulting model was then used to perform a cross validation analysis of the number of predictive variables that best predict the outcome of interest. It turns out that the top 8 variables ranked by importance reduce the error.

A second random forest with 800 trees was fit using only the 8 variables selected from the first result. Both random forests were compared by evaluating their robust mean squared prediction error using the evaluation data set.

3 Results

3.1 Linear model

The step-wise variable selection procedure determined that the variables *amount requested, amount funded by investors, loan length, debt to income ratio, issued date, inquiries in the last 6 months, issued date, earliest credit line, FICO (numerical), log monthly income, loan purpose, home ownership status, and log revolving credit balance* were important. Furthermore, this model has good model diagnostics and is significantly better than the naive model (p-value less than 2e-16). The full data model did not perform better than the variable selected model.

3.2 Random Forest

The cross validation of feature selection on the first random forest determined that that the it performs the best when using the top 8 variables ranked by importance. These are *FICO score (numerical), loan length, amount funded by investors, amount requested, issued date, inquiries in*

the last 6 months, open credit lines, and log revolving credit balance. This is a subset of the variables selected by the step-wise variable selection with the linear model.

When using the evaluation data set to calculate the robust mean squared prediction error, the model that performed the best is the random forest using the 8 variables described previously as shown in Table **??**.

	RMSPE	SE	Model
1	1.98	0.06	LM step
2	1.59	0.06	Initial Rand. Forest
3	1.55	0.06	Rand. Forest with top 8 vars.

Table 1: Robust mean squared prediction error for the step-wise variable selected model from linear regression, the initial random forest, and the random forest with the top 8 variables ranked by importance. Numbers are rounded to two digits.

4 Conclusions

There are 7 variables that can help explain the interest rate once the FICO score has been taken into account. These were determined by using random forests trained on 70% of the data. The resulting model has a robust mean squared prediction error of 1.554 with an standard error of 0.05611.

This model outperformed a naive initial random forest using all the variables (after pre and post processing) and a step wise variable selection procedure on a multiple linear regression model. The difference is rather significant with the latter of the two.

This analysis is reproducible in its entirety by using the *lcollado754* package. This includes the exploratory data analysis and variable selection steps that can be reproduced as HTML reports using *knitr*. Furthermore, this report is the vignette from the package and can be re-built by the user if such a thing is desired.

5 Reproducibility

Please check `https://github.com/lcolladotor/lcollado754/tree/master/final` for details on how to install the *lcollado754* package and reproduce the results.

6 References

1. Downloaded the states regions from `https://aqs.epa.gov/aqsweb/codes/data/StateCountyCodes.csv`.

2. The *knitr* package `http://yihui.name/knitr/` was heavily used in this analysis.

3. HTML reports are formatted for maximum cool-ness factor using Knitr Bootstrap `https://github.com/jimhester/knitr_bootstrap`.