Credit Risk via Lending Club

Modern Data Mining

Contents

Instructions

- This is a project. Well organized and well presented write-up is one major motivation here. Please see the section on `Write up` for details.
- There is no single correct answer.

- The entire write up should not be more than **5** pages. All the R-codes should be hidden. Any R-output used should be formatted neatly. You may put all supporting documents, graphics, or other exhibits into an Appendix, which is not counted in the 5 page limit.

Introduction

Background

Lending Club https://www.lendingclub.com/ is the world's largest peer-to-peer online platform connecting borrowers and investors. By cutting off the middle financial institutions in traditional lending process, Lending Club provides higher return for individual investors and lower interest rates for borrowers to get access to funding.

Treasury bonds are always considered to be risk-free since they are backed by the U.S. government. However, the interest rate of 10-year T-bond decreased from 4.68% in 2007 to 1.97% in 2015, and many investors are trying to find better alternative choices for investment. Lending Club offers an attractive choice. As such, Lending Club's business has grown exponentially in the loan market in recent years, so it's more and more important for Lending Club to control risks by distinguishing good (not default, borrowers pay back loans monthly) and bad (default, borrowers do not pay back) loans.

In the dataset provided, only the applications that were approved by Lending Club and provided a loan are included. This means that Lending Club has already filtered out some of the applicants based on certain criteria. We will restrict our analysis to the period between 2007-2011 for which we have around 39,000 observations and 38 attributes for each of these loans. These attributes include loan amount, home ownership status, interest rate on the loan, loan status and grade of the loan among many others.

Data needed:

- **LoanStats_07_11_Full.csv**
- **LoanStats_07_11_Clean.csv**

Goals of the study

1. You are hired by a large investment firm to invest their money through Lending Club. You are going to apply the machine learning skills from our data mining class and recommend a classification rule to identify types of loans to be included in the portfolio. In particular, you need to provide the following information

i) First, identify the important risk factors that a loan will be defaulted.

- This will require you to report a model to characterize the relationship between risk factors to the chance of a loan being defaulted.

- The set of available predictors is not limited to the raw variables in the data set. You may engineer any factors using the data that you think will improve your model's quality.

ii) Build a classifier that maximizes the return.

- To do so you need to propose a sensible loss ratio for picking up a bad loan. We did a quick estimate that the loss ratio of picking up a bad loan to that of missing a good loan is about 2 to 1. You may modify this loss ratio with your reasoning.

- You may build a few models possibly using elastic net, direct logistic regression etc. and choose among them. Your final classifier built should have a good testing power. I.e., it should work well for a testing data set that you may reserve from the data provided.

- Notice that the model behind the classifier you build here may differ from that constructed in question i).

2. You are young and ambitious. You see the opportunity of getting into this business. Based on the information available from the Lending club website and the analyses you have done, summarize why the lending club is so successful and has been able to grow the business rapidly. To do so you may need to gather further information to better understand Lending Club's business model.

3. Based on your knowledge gathered so far and analyses you have done, what can you offer to the Lending Club so that they can modify their selection rules to increase the returns for investors. [You may need go beyond the dataset provided.]

4. You may propose your own goal of study and ignore the agenda proposed above!

5. We suggest you to split the data first to Training/Testing/Validation data:

- Use training/testing data to land a final model (If you only use LASSO to land a final model, we will not need testing data since all the decisions are made with cross-validations.)

- Evaluate the final model with the validation data to give an honest assessment of your final model.

Characteristics of the Data Set

The data ranges from 2007 to 2011. We do not include data after 2012 because there might be loans that have not been closed. The original dataset is from https://www.lendingclub.com/info/download-data.action (sorry no longer available) and we drop irrelevant variables and variables with lots of missing values.

Description of variables

The variables could broadly be segmented into pre-funded loan data, borrower data, borrower credit data and post-loan data.

a) Pre-funded loan data a.`loan_amnt`: The listed amount of the loan applied for by the borrower b. `int_rate`: Interest Rate on the loan c. `grade`: LC assigned loan grade d. `sub_grade` : LC assigned loan subgrade e. `installment`: The monthly payment owed by the borrower if the loan originates f. `purpose`: The monthly payment owed by the borrower if the loan originates. g. `term`: The number of payments on the loan. Values are in months and can be either 36 or 60.

Remark: LC Grade

Loans **grade** tranches "credit-worthy" borrowers into seven investment grades, each with five subgrades (for a total of 35 tranches). Lending Club's methodology for tranching its borrowers follows proprietary scoring models based on FICO scores, credit history, "certain other credit attributes", loan term and loan amount.

Investors can decide how much to fund each borrower (subject to a minimum investment amount of $25/loan) and the proportion to invest in different loans of grade.

For instance, a loan in tranche A, the highest-rated tranche may default while a loan in trache E, the lowest-rated tranche, may pay in full. To be sure, variance is expected. No model is perfect. However, the variance made us realize that there were some assets that were priced inefficiently and thus, there existed an abritrage opportunity. If we could more accurately identify the loans that are likely to default and those that are likely to pay in full, then we could profit disproportionately.

Loan	Grade Interest	Rate Origination	Fee 36-Month	APR 60-Month
A	5.32% - 7.97%	1% - 5%	5.99% - 11.49%	7.46% - 10.17%
B	9.44% - 11.99%	5%	12.99% - 15.59%	11.67% - 14.27%
C	12.62% - 16.02%	6%	16.99% - 20.49%	15.40% - 18.89%
D	17.09% - 21.45%	6%	21.59% - 26.07%	19.99% - 24.48%
E	19.99% - 26.30%	6%	24.57% - 31.06%	22.98% - 29.49%
F	24.24% - 30.75%	6%	28.94% - 35.64%	27.36% - 34.09%
G	28.55% - 30.99%	6%	33.37% - 35.89%	31.82% - 34.34%

b) Borrower basic information a. `emp_title`: The job title supplied by the Borrower when applying for the loan b. `emp_length`: Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years. c. `home_ownership`: The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER. d. `annual_inc`: The self-reported annual income provided by the borrower during registration e. `zip_code`: The first 3 numbers of the zip code provided by the borrower in the loan application f. `addr_state`: The state provided by the borrower in the loan application g. `verification_status`: Indicates if income was verified by LC, not verified, or if the income source was verified

c) Borrower credit data a. `dti`: A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income. b. `delinq_2yrs`: The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years c. `earliest_cr_line`: The month the borrower's earliest reported credit line was opened d. `inq_last_6mths`: The number of inquiries in past 6 months (excluding auto and mortgage inquiries) e. `open_acc`: The number of open credit lines in the borrower's credit file. f. `pub_rec`: Number of derogatory public records g. `revol_bal`: Total credit revolving balance h. `revol_util`: Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit. i. `total_acc`: The total number of credit lines currently in the borrower's credit file j. `pub_rec_bankruptcies`: Number of public record bankruptcies

Credit data of the borrower, such as their debt to income ratio, delinquencies, credit inquiries, etc.

d) Post-loan data a. `issue_d`: The month which the loan was funded b. `loan_status`: Current status of the loan c. `funded_amnt`: The total amount committed to that loan at that point in time. d. `funded_amnt_inv`: The total amount committed by investors for that loan at that point in time e. `total_pymnt`: Payments received to date for total amount funded f. `total_pymnt_inv`: Payments received to date for portion of total amount funded by investors g. `total_rec_prncp`: Principal received to date h. `total_rec_int`: Interest received to date i. `total_rec_late_fee`: Late fees received to date j. `recoveries`: post charge off gross recovery k. `collection_recovery_fee`: post charge off collection fee l. `last_pymnt_d`: Last month payment was received m. `last_pymnt_amnt`: Last total payment amount received n. `last_credit_pull_d`: The most recent month LC pulled credit for this loan

Including post-loan data is unrealistic to predict the default rate *before* making the investment!

Response: `loan_status`

loan_status	Description
Defaulted	Overdue for over 90 days
Charged off	defaulted and there is no longer a reasonable expectation of further payments (e.g. bankruptcy).

Fully Paid |

To save your time we are going to use some data sets cleaned by us. Thus, we provide two datasets:

`LoanStats_07_11_Full.csv` is the original data. You may use it for the purpose of summary if you wish. You will see that the original data can't be used directly for your analysis, yet.

`LoanStats_07_11_Clean.csv` is a cleaned version and they are modified in the following ways:

1) Columns with lots of NAs are excluded.
2) `pymnt_plan`, `out_prncp`, `out_prncp_inv`, `collections_12_mths_ex_med`, `chargeoff_within_12_mths`, `tax_liens`, `initial_list_status`, `application_type`, `policy_code` have little variability, and are as such excluded.
3) Drop `title` since it has similar explanatory value as `purpose`.
4) Drop `emp_title` because it has too many levels, but it is possible to classify them into different sectors.
5) Drop all rows with NAs directly after 1)-4) at the expense of 2% loss of data.
6) You might include `desc` back in the data to do text mining.

Suggested outline for your report

As you all know, it is very important to present your findings well. To achieve the best possible results you need to understand your audience.

Your target audience is a manager who holds an MBA, is familiar with financial terminology, and has gone through a similar course to our Modern Data Mining with someone like your professor. You can thus assume some level of technical familiarity, but should not let the paper be bogged down with code or other difficult to understand output.

Note then that the most important elements of your report are the clarity of your analysis and the quality of your proposals.

A suggested outline of the report would include the following components:

1) Executive Summary

- This section should be accessible by people with very little statistical background (avoid using technical words and no direct R output is allowed)
- Give a background of the study. You may check the original website or other sources to fill in some details, such as to why the questions we address here are important.
- A quick summary about the data.
- Methods used and the main findings.
- You may use clearly labelled and explained visualizations.
- Issues, concerns, limitations of the conclusions. This is an especially important section to be honest in - we might be Penn students, but we are statisticians today.

2) Detailed process of the analysis

i) Data Summary /EDA

- Nature of the data, origin
- Necessary quantitative and graphical summaries
- Are there any problems with the data?
- Which variables are considered as input

1.1 EDA First we read in the data.

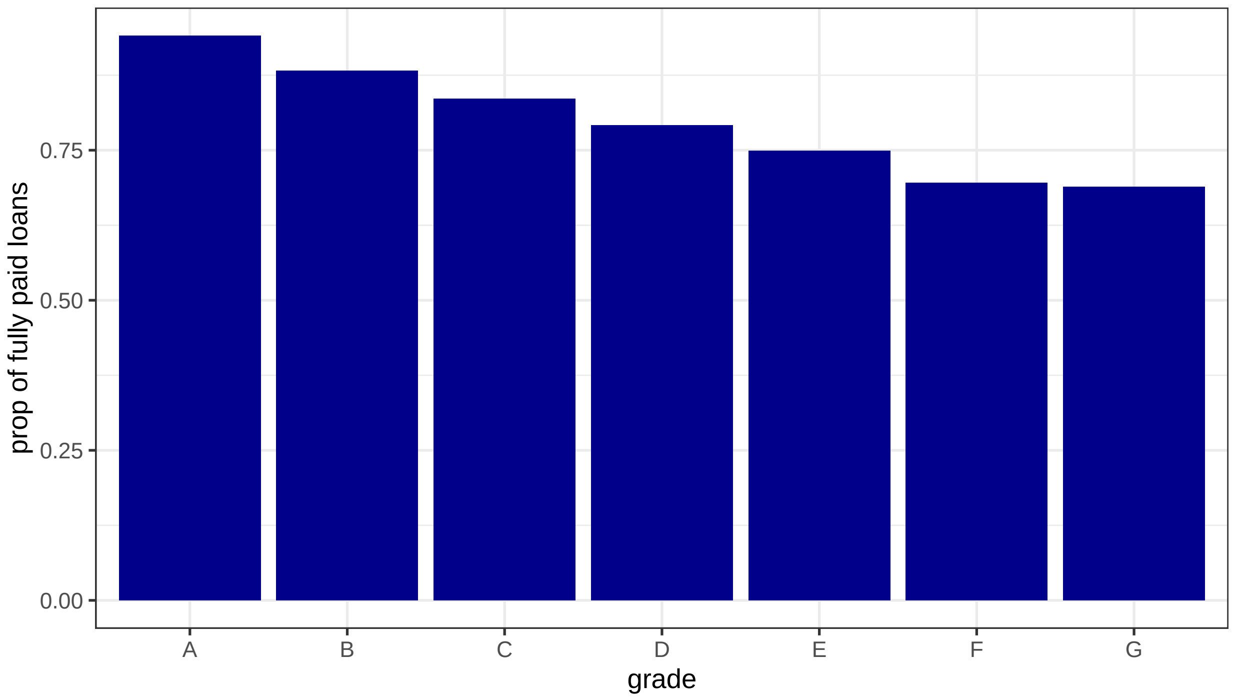
```
lending_data <- fread("LoanStats_07_11_Clean.csv")

view(dfSummary(lending_data),method = "render")
```

We coerces the loan_status column into type "factor". As shown in the barplot below, the percentage of fully-paid loans decreases steadily from Grade A to Grade G. Approximately 6% of Grade A loans defaulted whereas 31% of Grade G loans defaulted.

```
lending.sum = lending_data[order(grade),.(
  fully.paid.prop = length(which(loan_status == "Fully Paid"))/length(loan_status)
),by=grade]

ggplot(data = lending.sum,aes(x=grade,weight=fully.paid.prop))+
  geom_bar(fill = "Darkblue")+
  ylab("prop of fully paid loans")+
  theme_bw()
```



1.2 loan__status vs. annual__inc

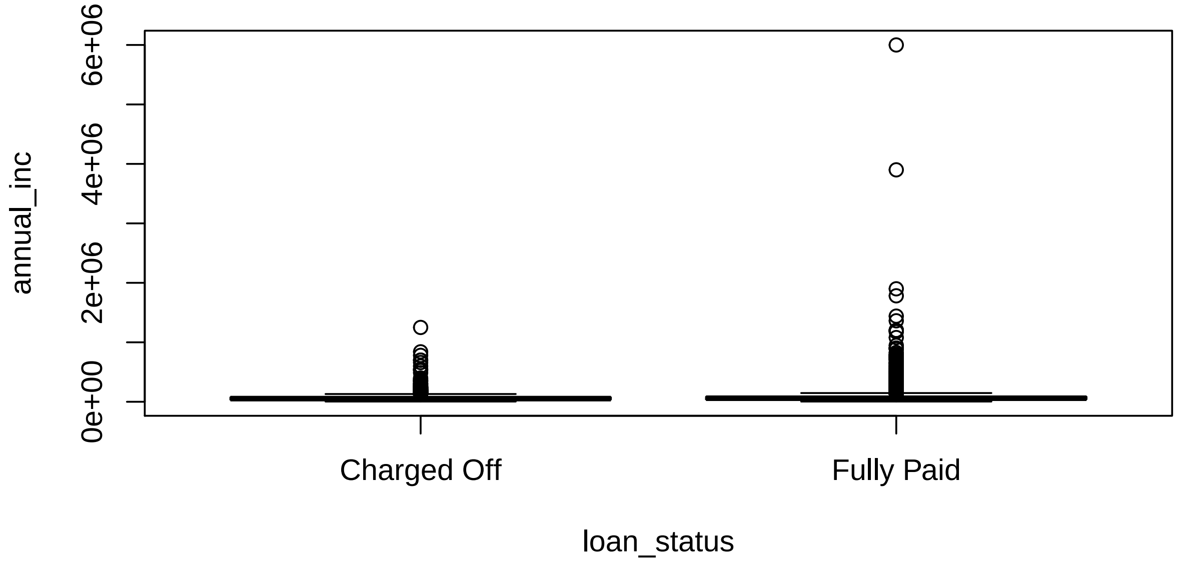
For simplicity we start with a simple question: How does loan status relate to the annual income of the individual

```
lending_data %>% group_by(loan_status) %>% summarise(mean(annual_inc))
```

```
## # A tibble: 2 x 2
##   loan_status `mean(annual_inc)`
##   <chr>                    <dbl>
## 1 Charged Off             62638.
## 2 Fully Paid              70082.
```

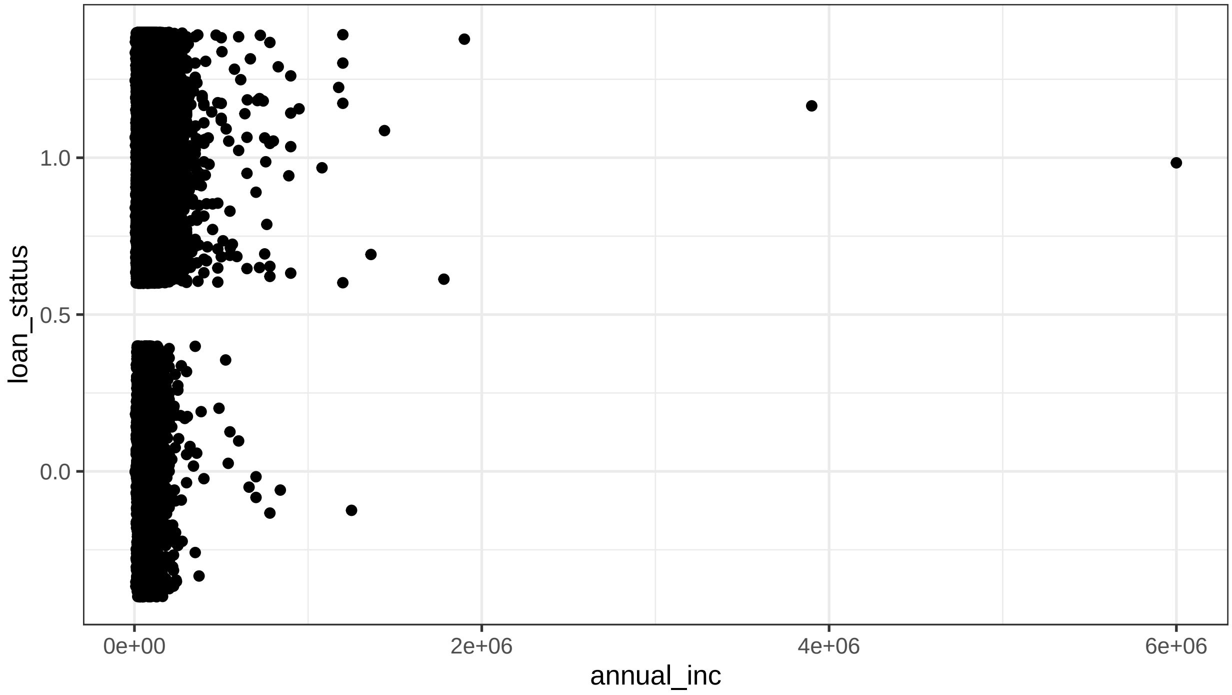
We observe on average, loans in default tend to be held by those with lower income.

```
boxplot(annual_inc~loan_status,lending_data)
```



Next we let Fully Paid = 1, and Charged Off = 0.

```
lending_data[,loan_status := as.numeric(loan_status == "Fully Paid")]
ggplot(data = lending_data,aes(x=annual_inc,y=loan_status))+
  geom_point(position = "jitter")+
  theme_bw()
```



2.1 Logistic Regression: loan_status vs. annual_inc

```
fit1 <- glm(loan_status~annual_inc, lending_data, family=binomial(logit))
summary(fit1, results=TRUE)
```

```
##
## Call:
## glm(formula = loan_status ~ annual_inc, family = binomial(logit),
##     data = lending_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6109   0.5085   0.5506   0.5713   0.6156
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.551e+00  2.914e-02  53.237   <2e-16 ***
## annual_inc  3.973e-06  3.985e-07   9.971   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 31607  on 38970  degrees of freedom
## Residual deviance: 31489  on 38969  degrees of freedom
## AIC: 31493
##
## Number of Fisher Scoring iterations: 5
```

```
fit1 <- glm(loan_status~annual_inc, lending_data, family=binomial(logit))
summary(fit1, results=TRUE)
```

```
##
## Call:
## glm(formula = loan_status ~ annual_inc, family = binomial(logit),
##     data = lending_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6109   0.5085   0.5506   0.5713   0.6156
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.551e+00  2.914e-02  53.237   <2e-16 ***
## annual_inc  3.973e-06  3.985e-07   9.971   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 31607  on 38970  degrees of freedom
## Residual deviance: 31489  on 38969  degrees of freedom
## AIC: 31493
##
## Number of Fisher Scoring iterations: 5
```

2.2 Inference for the Coefficients 2.2.1 Wald intervals/tests (through the MLE's)

```
confint.default(fit1)
```

```
##                     2.5 %       97.5 %
## (Intercept) 1.494324e+00 1.608558e+00
## annual_inc  3.192224e-06 4.754238e-06
```

2.2.2 Likelihood Ratio Tests

```
anova(fit1, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: loan_status
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       38970      31607
## annual_inc  1   118.75     38969      31489 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3 Multiple Logistic Regression and Classification

4 Decision Tree

Cleaning the Data

```
tree_data <- lending_data %>% select(c(loan_status, dti, delinq_2yrs,  inq_last_6mths, revol_util, revol
tree_data$loan_status <- as.factor(tree_data$loan_status)
```

Create a function that splits data into train and test sets

```
create_train_test <- function(data, size = 0.8, train = TRUE) {
    n_row = nrow(data)
    total_row = size * n_row
    train_sample <- 1: total_row
    if (train == TRUE) {
        return (data[train_sample, ])
    } else {
        return (data[-train_sample, ])
    }
}
```

Split the Data into train/test set

```
data_train <- create_train_test(tree_data, 0.8, train = TRUE)
data_test <- create_train_test(tree_data, 0.8, train = FALSE)
dim(data_train)
```

```
## [1] 31176    10
```

Verify that the randomization process is correct

```
prop.table(table(data_train$loan_status))
```

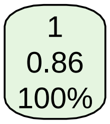
```
##
```

```
##         0         1
## 0.1446626 0.8553374
```

```
prop.table(table(data_test$loan_status))
```

```
##
##         0         1
## 0.1228993 0.8771007
```

```
fit_tree <- rpart(loan_status~., data = data_train, method = "class", minsplit = 2, maxdepth=8)
rpart.plot(fit_tree)
```



ii) Analyses

- Various appropriate statistical methods: e.g. glmnet (and/or trees, ignore this at the moment)
- Comparisons various models
- Final model(s)

iii) Conclusion

- Summarize results and the final model
- Caveats
- Final recommendations

Maintain a good descriptive flow in the text of your report. Use Appendices to display lengthy output.

iii) Appendix

- All your R code (code without comments is no good!) if you are not using `rmd` format.
- Any thing necessary to keep but for which you don't want them to be in the main report.