Loan Performance

Lending Club
Capstone Project Report
15-Jun-2016

Name
Madhu Samudrala

Acknowledgements

I would like to take this opportunity to express my profound gratitude and deep regards to our project mentor, Professor Minta Thomas for her guidance throughout the course of this project work.

I am deeply indebted to all the Professors who taught us various courses during Business Analytics program.

Madhu Samudrala

Contents

Overview

Lending industry is a major financial sector. Companies in this industry provide secured and unsecured loans and other credit and financing products. Demand is driven by interest rates, consumer confidence, and capital spending by businesses. The profitability of individual companies depends on their ability to originate, service, and collect loans, as well as to collect fees and interest on credit and other financing products.

There is an increasing use of technology in identifying patterns of consumer payment behavior and how it will impact business.

The main challenge faced by most financial institutions is determining probability of default. If a lender could forecast how many loans would be paid by the term, how many would become delinquent and how many loans would be re-structured, it would help the lender plan investment sources better, manage cash flows efficiently.

Analytics will help solve the challenge by using various forecasting and predictive models implemented over huge datasets of historical data.

About Lending Club

Lending Club is a US peer-to-peer lending company, headquartered in San Francisco, California, has been operating for the past nine years.

Lending Club conducts its operations completely online without any branch infrastructure. The no branch model helps in saving costs which is passed on to borrowers as lower interest when compared to other conventional lending agencies.

They provide personal loans, business loans and loans for elective medical procedures. Borrowers access loans at a lower interest rates and investors pool in money to earn interest.

Problem Statement

- The main challenge faced by most financial institutions is determining probability of default.

- If a lender could forecast how many loans would be paid by the term, how many would become delinquent and how many loans would be re-structured, it would help the lender plan investment sources better, manage cash flows efficiently.

- This would also save a lot of cost on unnecessary administrative and legal procedures and collection efforts.

- This model is aimed to calculate probability of default, considering certain key data point to ease the problem described above.

About the Dataset

- The dataset is shared on the company’s website, available for everyone to access.
- These files are csv downloads that contain complete loan data.
- There is also a Data Dictionary available on the website that includes definitions for all the data attributes included in the data file.

Key Attributes

#	Attribute name	Description
1	loan_amnt	The listed amount of the loan applied for by the borrower.
2	int_rate	Interest Rate on the loan
3	Term	The number of payments on the loan. Values are in months and can be either 36 or 60.
4	Grade	LC assigned loan grade
5	home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
6	issue_d	The month which the loan was funded
7	loan_status	Current status of the loan
8	Purpose	A category provided by the borrower for the loan request.
9	addr_state	The state provided by the borrower in the loan application
10	total_pymnt	Payments received to date for total amount funded
11	inq_last_6mths	The number of inquiries by creditors during the past 6 months.
12	last_fico_range_high	The last upper boundary of range the borrower’s FICO belongs to pulled.
13	last_fico_range_low	The last lower boundary of range the borrower’s FICO belongs to pulled.

Number of Orders

- Total 1,007,217 record for years 2007 through 2015.

Challenges Faced

- I had a tough time deciding the technology and tools to use for the project. Finally after a lot of research decided to build the predictive model on R's random forest and a Tableau dashboard.
- Cleaning and processing the data in R was tough as I am new to programming. I did a lot of research on the internet to get my code working.
- Deploying the predictive model on Shinny app was the toughest challenge I faced in this project. I consulted friends, mentor, researched on the internet. Took me months to finally deploy the app.
- With the Lending Club dataset plotting filled maps on tableau was not possible. I had to use a secondary source to build a polygon map within tableau. The secondary source contains Point Order, Polygon ID, latitude and longitude. This helped in plotting the filled map.
- Plotting text labels on polygon chart was one of the challenge. I had to build a dual access map to get the text labels shown on the map.

Exploratory Data Analysis, Processing and Prediction

The publicly available dataset of Lending Club is a real world data set with a nice mix of categorical and continuous variables.

After a thorough study on the dataset, I decided to use the file from 2007 to 2011 to build the prediction model. The rationale behind this was to limit the build sample to mature vintages only. In other words, if we were to look at loans originated in 2012, some would only be part-way through repayment and therefore would appear to be performing better than mature vintages which have had more time to go bad.

There are many machine learning algorithms for predicting bad loans. Of all, Random Forest does a pretty outstanding job with most prediction problems, so I decided to use R's Random Forest package for this model.

Library’s used

- library(stringr)
- library(plyr)
- library(lubridate)
- library(randomForest)
- library(reshape2)
- library(caret)
- library(shiny)
- library(e1071)

Techniques Used

Below are the techniques used:

- Load data into R

```
df <- read.csv('LoanStats.csv', h=T, stringsAsFactors=F, skip=1)
```

- Get rid of NULL & N/A (annoying columns)

```
df[,'desc'] <- NULL

df[,'mths_since_last_record'] <- NULL
```

```
poor_coverage <- sapply(df, function(x) {

  coverage <- 1 - sum(is.na(x)) / length(x)

  coverage < 0.8

})

df <- df[,poor_coverage==FALSE]
```

- Define bad_indicators

```
bad_indicators <- c("Late (16-30 days)", "Late (31-120 days)", "Default", "Charged Off")



df$is_bad <- ifelse(df$loan_status %in% bad_indicators, 1,

                    ifelse(df$loan_status=="", NA,

                           0))
```

- Read and clean

```
table(df$loan_status)

table(df$is_bad)

df$issue_d <- as.Date(df$issue_d,format = "%m/%d/%Y")

df$year_issued <- year(df$issue_d)

df$month_issued <- month(df$issue_d)

df$earliest_cr_line <- as.Date(df$earliest_cr_line, format = "%m/%d/%Y")



df$revol_util <- str_replace_all(df$revol_util, "[%]", "")

df$revol_util <- as.numeric(df$revol_util)

outcomes <- ddply(df, .(year_issued, month_issued), function(x) {
 c("percent_bad"=sum(x$is_bad) / nrow(x),
   "n_loans"=nrow(x))
})




df.term <- subset(df, year_issued < 2012)

df.term$home_ownership <- factor(df.term$home_ownership)

df.term$is_rent <- df.term$home_ownership=="RENT"
```

○ Split the data

```
idx <- runif(nrow(df.term)) > 0.75

train <- df.term[idx==FALSE,]

testData <- df.term[idx==TRUE,]
```

○ Run RF

```
df$is_bad <- ifelse(df$is_bad == 1, "X", "Y")

fitControl <- trainControl (method = "cv", number = 3)

rfFit <- train (factor(is_bad) ~ last_fico_range_high + last_fico_range_low +
                  revol_util + inq_last_6mths,
                data=df[1:100,c('is_bad','last_fico_range_high','last_fico_range_low',
                         'revol_util','inq_last_6mths')],method = "rf",
                trControl = fitControl, verbose = FALSE)
```

- **Deploying to Shiny:**

  I've got this script, but these insights would be more useful in a live application. So I've deployed this on Shiny app.

  - server.R

  the above code +

```
shinyServer(
  function(input, output, session){


    output$loandf = renderText ({


      min1<- input$Min_FICO_Score
      max1<- input$Max_FICO_Score
      rev<- input$Revolving_Line_Utilization
      cred<- input$Credit_Inquiries_Past_6m
      hom<- input$Home_Ownership
      ann<- input$Annual_Income
      loan<- input$Loan_Amount



      testData$last_fico_range_high <- as.numeric(min1)
```

```
   testData$last_fico_range_low<-as.numeric(max1)
   testData$revol_util <- as.numeric(rev)
   testData$inq_last_6mths <- as.numeric(cred)
   testData$home_ownership<-as.character(hom)
   testData$annual_inc<-as.numeric(ann)
   testData$loan_amnt<-as.numeric(loan)



   summary(df[1,])

   round(predict (rfFit, newdata=as.data.frame(testData[1,]),type='prob')[,1],3)

  })

 })
```

- ui.R

```
shinyServer(
 pageWithSidebar(
  headerPanel("Predict Bad Loan Applicant"),

  sidebarPanel(
      textInput("Min_FICO_Score","Min FICO Score",500),

      textInput("Max_FICO_Score","Max FICO Score",600),

      textInput("Revolving_Line_Utilization","Revolving Line Utilization",20),

      textInput("Credit_Inquiries_Past_6m","Credit Inquiries Past 6m",1),

      selectInput("Home_Ownership","Home Ownership", choices=c("RENT",
"OWN","MORTGAGE")),

      textInput("Annual_Income","Annual_Income",75000),

      textInput("Loan_Amount","Loan_Amount",6000)

    ),

    mainPanel(

      h2 ('Probability of default'),

      h3 (textOutput ('loandf')),

      tags$style ("#loandf{color: red;

            font-size: 25px;

            font-style: bold;

            }"

       ),


      #h2 ('You Entered'),


        tags$div(class="modal-footer",

            "Note: This is a capstone course project by Madhu Samudrala. The dataset used for
prediction is 'Lending Club Loan Dataset' and the
```

```
          prediction algorithm is Random Forest.")

     )



))
```

URL: https://madhusamudrala.shinyapps.io/loan/

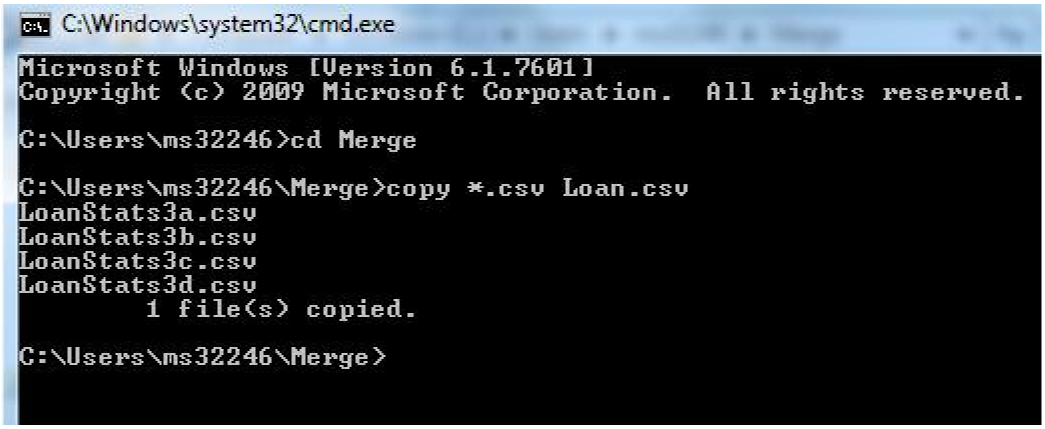
Visualization and Analytics

Data visualization and analytics is the process of describing information through visual rendering. Humans have used visualizations to explain the world around them for millions of years. Data visualization allows for universal and immediate insight by tapping into our mind's powerful visual processing system.

Tableau leads the world in making the data visualization process available to business users. Tableau Software enables businesses to keep pace with the evolving technology landscape and outperform competitors through an adaptive and intuitive means of visualizing their data. Hence I decided to use this tool for my project.

Load data to Tableau

- The downloaded multiple csv files from Lending Club website were merged using command prompt.



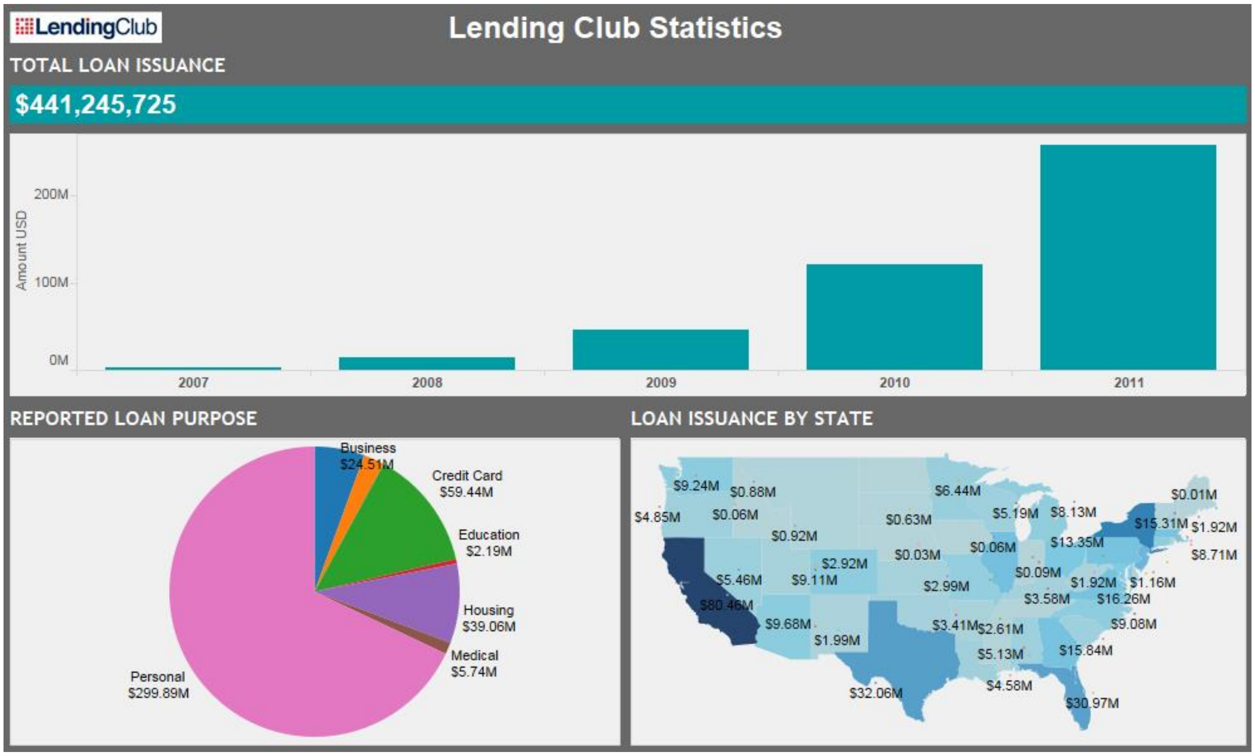
```
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation.  All rights reserved.

C:\Users\ms32246>cd Merge

C:\Users\ms32246\Merge>copy *.csv Loan.csv
LoanStats3a.csv
LoanStats3b.csv
LoanStats3c.csv
LoanStats3d.csv
        1 file(s) copied.

C:\Users\ms32246\Merge>
```

- Downloaded 'State Polygons.xlsx' file (url: http://kb.tableau.com/articles/knowledgebase/polygon-shaded-maps) . This file is used as secondary source to build polygon maps within Tableau.

Dashboards

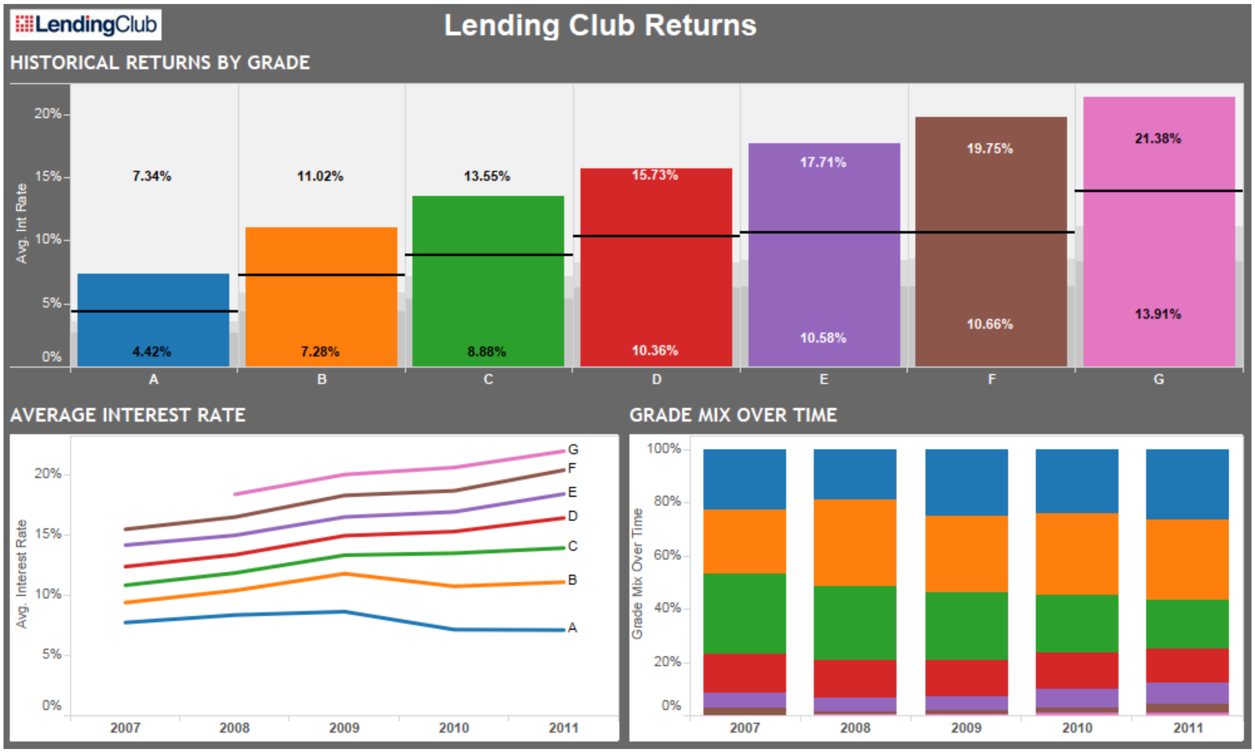
Lending Club Statistics:

- This dashboard shows companies performance from inception, till 2015.
- The number on the top left is the total amount funded.
- The bar chart is the break of the amount spread across multiple years.
- The pie depicts amount funded for various loan types.
- The map depicts amount funded in each state.



Lending Club Returns:

- This dashboard shows the returns earned/received by the financial institution.
- The chart 'HISTORICAL RETURNS BY GRADE' shows average lending interest rate and the interest rate earned for each Grade.
- The 'AVERAGE INTEREST RATE' chart shows the average interest rate for each Grade spread across multiple years.
- The 'GRADE MIX OVER TIME' chart depicts the percentage of loans lent for each Grade by year.



Lending Club Returns

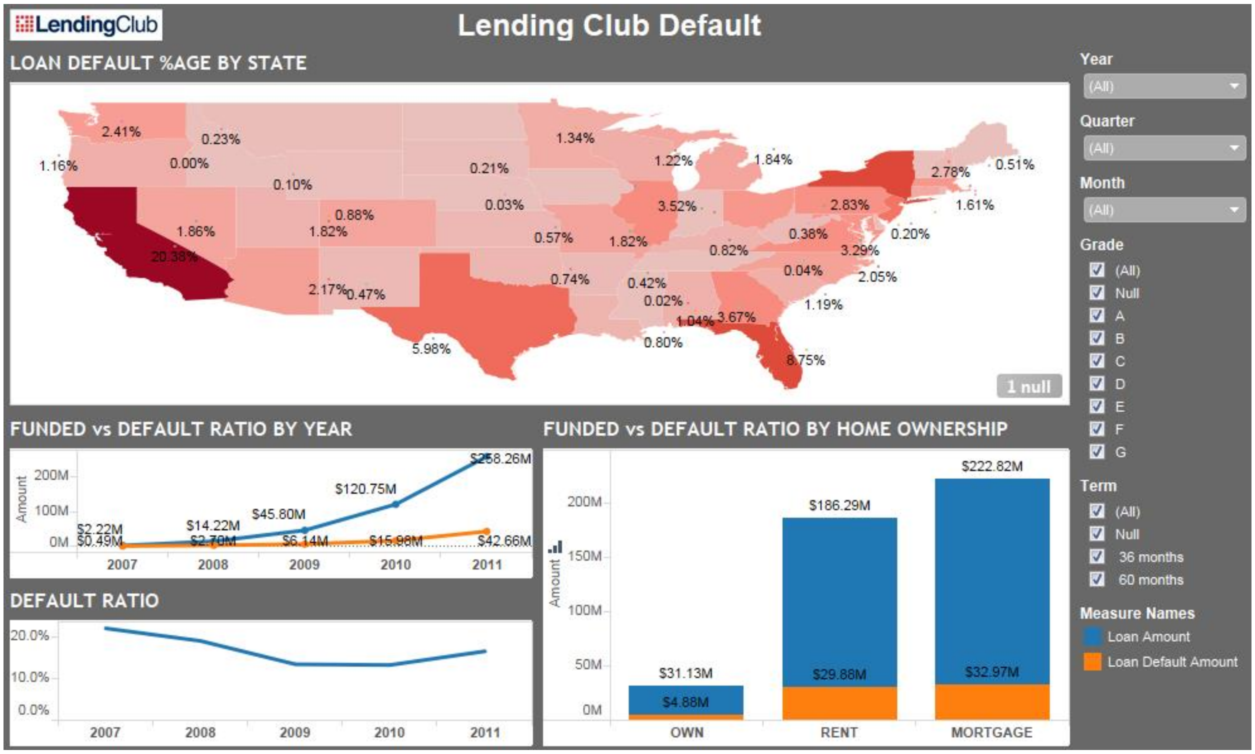
HISTORICAL RETURNS BY GRADE

AVERAGE INTEREST RATE

GRADE MIX OVER TIME

Lending Club Default:

- This dashboard shows default/loss incurred by the financial institution.
- The 'LOAN DEFAULT %AGE BY STATE' chart shows the percentage of defaulters across the country. The color coding in red, from dark to light, depicts that dark shade of red is where the most offenders are and the light shade is where there are less offenders.
- The 'FUNDED vs DEFAULT RATIO BY YEAR' chart shows the amount funded (blue line) each year and default amount (amber line) each year.
- The 'DEFAULT RATIO' trend line depicts the percentage of defaulters across each year.
- The 'FUNDED vs DEFAULT RATIO BY HOME OWNERSHIP' chart shows amount funded vs default amount by home ownership status.
- The dashboard can be further sliced/diced across period, grade, and term.



Lending Club Default

LOAN DEFAULT %AGE BY STATE

FUNDED vs DEFAULT RATIO BY YEAR

DEFAULT RATIO

FUNDED vs DEFAULT RATIO BY HOME OWNERSHIP

Year: (All)

Quarter: (All)

Month: (All)

Grade

☑ (All)
☑ Null
☑ A
☑ B
☑ C
☑ D
☑ E
☑ F
☑ G

Term

☑ (All)
☑ Null
☑ 36 months
☑ 60 months

Measure Names

🟦 Loan Amount
🟧 Loan Default Amount

*Please note that screen prints shown for 'Dashboards' segment are data from 2007 till 2011. The updated data (2007 to 2015) can be seen in Tableau Public post 18th June 2016 (url provided below).

Conclusion

I started off with a major challenge of developing a predictive/forecasting model to help make financial institutions better operational decisions.

This model helps in predicting probability of loan default.

I have leveraged data provided by Lending Club to build a Random Forest model powered by ‘R’ and deployed the same over the web.

For analysis, I have also built an interactive Tableau dashboard to visualize the data. The data from this dashboard can be sliced/diced as per user liking for further drill downs.

This project framework will give confidence to lending institutions to use forecasting model while making decisions and continuously enhance with more datasets.

References

- Dataset - https://www.lendingclub.com/info/download-data.action
- Predictive model - https://madhusamudrala.shinyapps.io/loan/
- Tableau Dashboard - https://public.tableau.com/views/LoanPerformance/LendingClubStatistics?:embed=y&:display_count=yes&:showTabs=y