Lending Club Capstone Final

Minh Truong

October 1, 2016

Table of Contents

## Lending Club (LC)

Lending Club is a new Peer2Peer Platform that connects lenders with borrowers (think Kickstarter for personal loans). LC uses an algorithm model to decide on the risk level or interest rate borrowers need to be charged lenders. Rates are made to be easy to apply and lending club charges a fee to finance for borrowers and charges lender's fees for the services rendered.

*PROBLEM:*

<http://www.bloomberg.com/news/features/2016-08-18/how-lending-club-s-biggest-fanboy-uncovered-shady-loans>

*Recently there has been a scandal where some borrowers may be the same borrowers and some anomalies were found.* Stock price has lost 2/3 of its value $15 to $5 almost overnight.

*TLDR; BAD LOANS ARE AFFECTING the STOCK PRICE AND PROFITABILITY LENDING CLUB*

*ISSUES/CHALLENGES:*

1. Can bad loans be identified before being accepted into the system?
2. What are the characteristics of a bad loan? Are there tell-tail signs?
3. What are the short/long term strategies after successful identifying?
4. What are other potential areas to improve or look at?
5. Are there are interesting or weird issues?

*FINDINGS:*

1. YES it is possible to make a bad loans model that can identify bad loans.
2. We are looking for any patterns, unusual clumping or clues on why it won't look like a normal healthy population distribution.
3. Implement ways to mitigate short term risks with accelerated payment inventives with longer term strategy to uncover hidden risks that look safe.
4. Can look at ways to run a linear regression on the top variables to improve bad loans model
5. Originally I was expecting the types of variables that would signal being unable to pay like credit inquiries, delinquencies, income, interest rates but what I found were variables that seem to signify people who would be a great credit risk but but had hidden circumstances that may or may not show up in the data directly and if they showed it would be indirect proof.

\*[DATA:\*](DATA:*)

URL: <https://www.lendingclub.com/info/download-data.action>

The dataset is HUGE 421k loans and 115 possible variables/features not counting features we can add

Since it is very time consuming to try all the combinations and nor do we want to build a model with too many variables as it will make the model unnecessarily complicated and the danger of overfitting the data.

Issue with overfitting would mean it explains the current data set but has poor predictability for new datasets.

Also since we have many observations and many variables the strategy will best to take advantage of what is available and try to extract as much value thats existing before trying anything too fancy.

STARTING WITH THE USUAL CUPRITS THAT MAY FIT (Who are the ones that we need to get?)

*Keep in mind most households have 15K in credit card debt, with many with auto loans, mortgages, student loans and other debt*

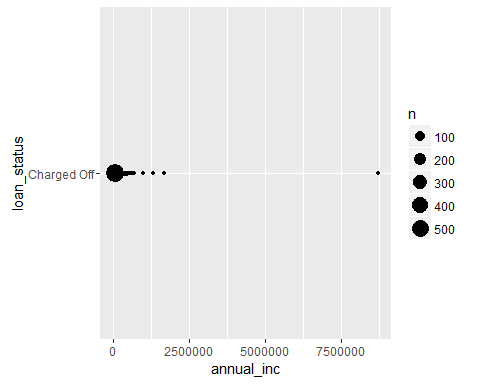
So It isn't unusual to borrow money but the clue is to if certain characteristics are overweighted that shouldnt be.

Some easy ones that come to mind are low income type loans, or high interest type loans, or people who carry an unusual balance loads/acconts or flags like credit inquires or past delinquencies that signal I may not pay.

suppressWarnings(suppressMessages(library(ggplot2)))  
suppressWarnings(suppressMessages(library(dplyr)))  
suppressWarnings(suppressMessages(library(tidyr)))  
suppressWarnings(suppressMessages(library(lubridate)))  
suppressWarnings(suppressMessages(library(caTools)))  
suppressWarnings(suppressMessages(library(ROCR)))  
suppressWarnings(suppressMessages(library(pROC)))  
  
LC <- read.csv("LoanStats3d\_securev1.csv")  
summary(LC$loan\_status)

## Charged Off Current Default   
## 13325 336267 539   
## Fully Paid In Grace Period Late (16-30 days)   
## 55905 3974 2515   
## Late (31-120 days)   
## 8570

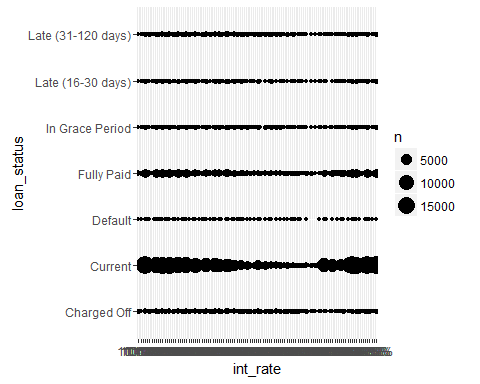
#Top reason many will say why they couldnt pay is "I don't make enough money", but #looking at the data set, there doesnt seem to be any huge defaults on the low end nor #the high end. Looking at the distribution quartiles, they look like income low or high #they default/charge off equally as likely, next most likely may be the interest rate, #"I got charged too much! Lets take a look."\*  
  
annual <- subset(LC,select = c("loan\_status","annual\_inc") )  
annual\_2 <- subset(annual, loan\_status == "Charged Off")  
ggplot(annual\_2, aes (x = annual\_inc, y =loan\_status ))+geom\_count()



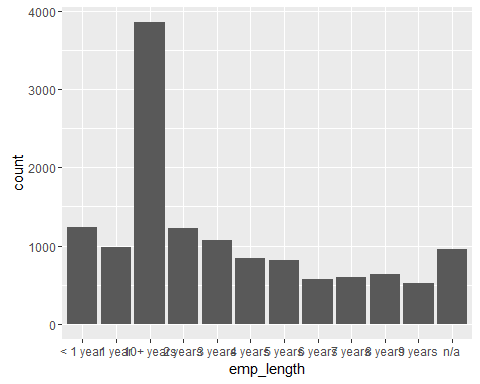
summary(annual\_2)

## loan\_status annual\_inc   
## Charged Off :13325 Min. : 0   
## Current : 0 1st Qu.: 42000   
## Default : 0 Median : 60000   
## Fully Paid : 0 Mean : 69714   
## In Grace Period : 0 3rd Qu.: 82800   
## Late (16-30 days) : 0 Max. :8706582   
## Late (31-120 days): 0

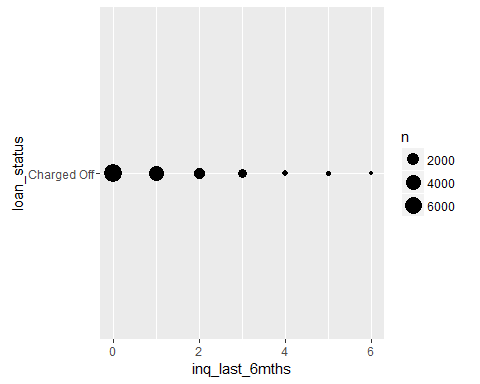
#Interest Rate seems like a good starting point, but doesn't seem to be a pattern, #there doesn't seem to be any unusual weights anywhere and seems fairly distributed #around. Lets see if having a job is the reason!  
   
int\_rate.loan\_status <- subset(LC,select = c("loan\_status","int\_rate") )  
ggplot(int\_rate.loan\_status, aes( x= int\_rate, y = loan\_status))+ geom\_count()



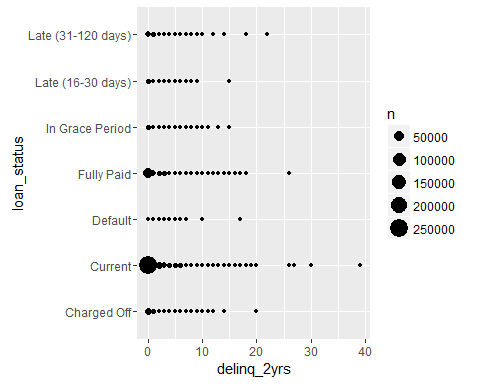
#Looks like employment length have alot of defaults with those @ 10+ Years, which is #unusual as those who are employed the longest tend to be responsible and consistent #or else they wouldnt be employed for so long. So this may be a clue on why. Lets see #if there are any delinquencies or credit inquiries.  
   
emp\_length <- subset(LC,select= c("loan\_status","emp\_length") )  
emp\_length\_3 <- subset(emp\_length, loan\_status == "Charged Off")  
emp\_length\_3 <- subset(emp\_length, loan\_status == "Charged Off")  
ggplot(emp\_length\_3,aes(x= emp\_length )) + geom\_bar()



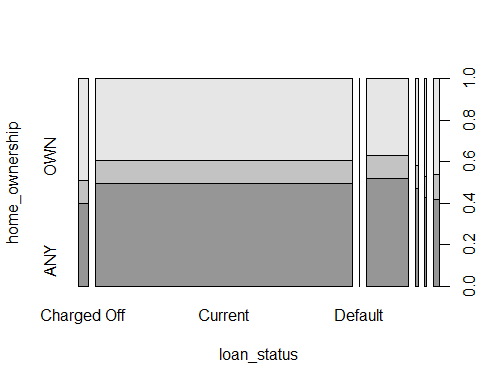
# Looks like an unusual of charge offs at 0 or 1 inquiries meaning most of the people #had few credit card binge opening, maybe they didnt open any new credit lines but #started to fall behind the ones they already have open. Lets find out!  
  
inq\_6mths <- subset(LC,select = c("loan\_status","inq\_last\_6mths") )  
inq\_6mths\_2 <- subset(inq\_6mths, loan\_status == "Charged Off")  
ggplot(inq\_6mths\_2, aes (x = inq\_last\_6mths, y =loan\_status ))+geom\_count()



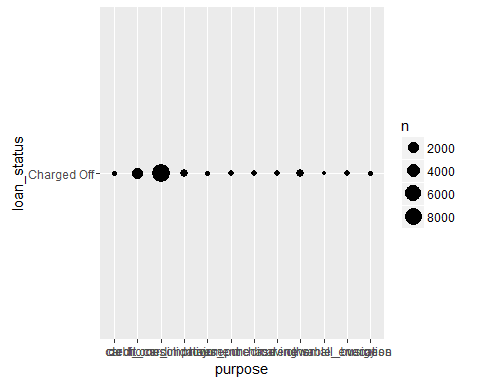
#So there's a very slight pattern of people with low to no delinquencies but very minor # you can argue it is no pattern as its so slight as many pay off their loans have no   
# delinquencies but can be a potential wolf in sheets clothing. Maybe its the housing #costs as typically rent or living costs take up to 30-50% of our income!  
  
delinq <- subset(LC,select = c("loan\_status","delinq\_2yrs") )  
ggplot(delinq, aes (x = delinq\_2yrs, y =loan\_status ))+geom\_count()



# See no trends here for homeownership, rent or own or in current mortgage. Maybe its #the reason why they borrowed!  
   
homeown <- subset(LC,select = c("loan\_status","home\_ownership") )  
 plot(homeown)



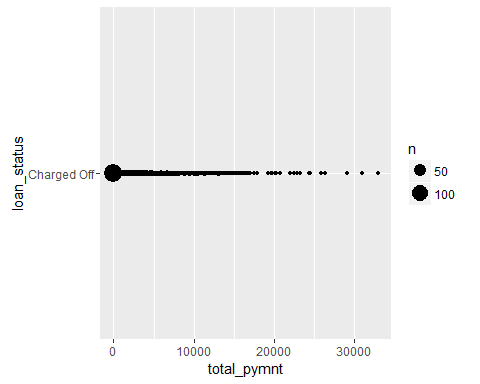
# Looking here I see a ton of credit card and debt consolidation as the biggest reasons # why people are borrowing.  
  
 purpose <- subset(LC,select = c("loan\_status","purpose") )  
purpose\_2 <- subset(purpose, loan\_status == "Charged Off")  
ggplot(purpose\_2, aes (x = purpose, y =loan\_status ))+geom\_count()



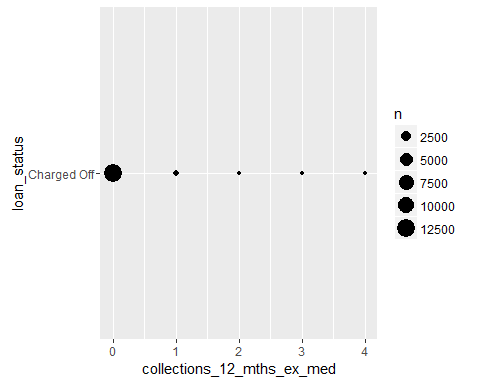
summary(purpose\_2)

## loan\_status purpose   
## Charged Off :13325 debt\_consolidation:8625   
## Current : 0 credit\_card :2141   
## Default : 0 other : 744   
## Fully Paid : 0 home\_improvement : 708   
## In Grace Period : 0 major\_purchase : 280   
## Late (16-30 days) : 0 small\_business : 215   
## Late (31-120 days): 0 (Other) : 612

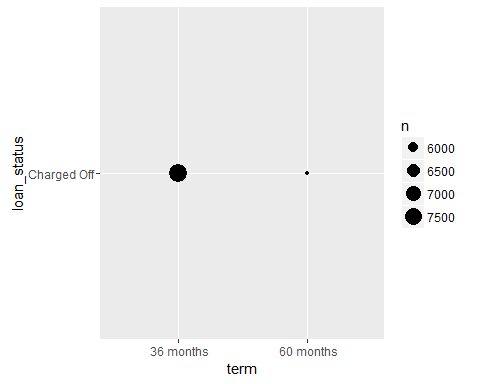
#So what have we learned so far?  
  
#Income/home ownership/ interest rates seem to have no patterns  
#Employment lengths/credit inquiried/delinquencies are long on the job but low on the #potential risk of default as none have triggered so far.  
  
#Seems to me so far all the data points to seemingly healthy profile digging a bit more #based on these characteristics, leads me to look for other areas of seemingly healthy #loan profiles like Payments!  
  
#the more payments made, less likely to charge off it looks like. So despite seemingly #healthy profiles people have trouble paying. Is it maybe they are sick? reckless?  
  
total\_payment <- subset(LC,select = c ("total\_pymnt","loan\_status") )  
total\_payment\_2 <- subset(total\_payment, loan\_status == "Charged Off" )  
ggplot(total\_payment\_2, aes(x= total\_pymnt, y= loan\_status) ) + geom\_count()



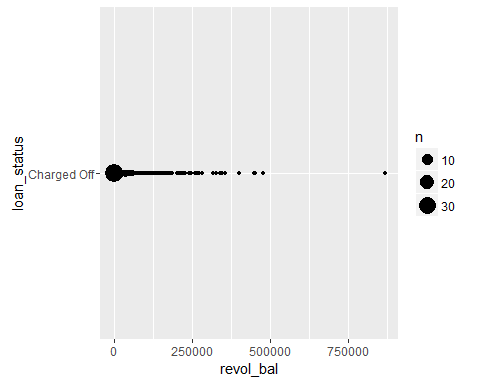
#medical collections seems most are healthy but things maybe happening behind closed #doors  
  
med\_coll\_12mth <- subset(LC, select = c("loan\_status","collections\_12\_mths\_ex\_med"))  
med\_coll\_12mth\_2 <- subset(med\_coll\_12mth, loan\_status == "Charged Off" )  
ggplot(med\_coll\_12mth\_2, aes(x= collections\_12\_mths\_ex\_med, y= loan\_status) ) + geom\_count()



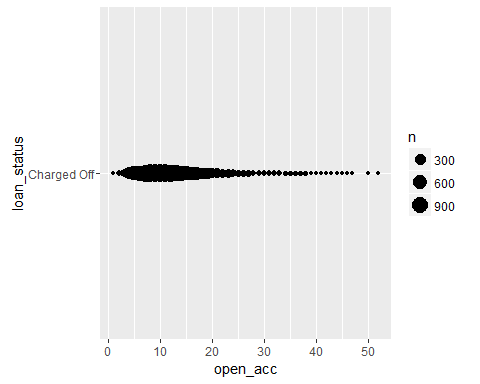
#a Rechlace person may borrow alot for as long as possible but most are short term #loans 36 months \*  
  
term <- subset(LC, select = c("loan\_status", "term"))  
term\_2 <- subset(term, loan\_status == "Charged Off")  
ggplot(term\_2, aes(x= term, y= loan\_status) ) + geom\_count()



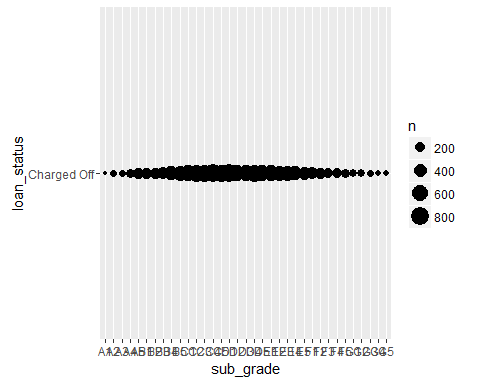
#revolv bal seems low too but this doesnt make sense why.  
  
rev\_bal <- subset(LC, select = c ("revol\_bal", "loan\_status"))  
rev\_bal\_2 <- subset(rev\_bal, loan\_status == "Charged Off")  
ggplot(rev\_bal\_2, aes(x= revol\_bal, y= loan\_status) ) + geom\_count()



#open accounts open arent unusually high either, it is healthy to have 5-20 credit #accounts open for a typical household between store cards, credit cards and other #likes like auto, student, mortgage  
  
open\_accounts <- subset(LC, select = c ("loan\_status", "open\_acc"))  
open\_accounts\_2 <- subset(open\_accounts, loan\_status == "Charged Off")  
ggplot(open\_accounts\_2, aes(x= open\_acc, y= loan\_status) ) + geom\_count()



#grade looks like a normal bell curve and not because LC allowed many high risk #profiles according to their algorithms  
  
grade <- subset(LC, select = c ("loan\_status", "sub\_grade"))  
grade\_2 <- subset(grade, loan\_status == "Charged Off" )  
ggplot(grade\_2, aes(x= sub\_grade, y= loan\_status) ) + geom\_count()

 WHAT DOES OUR BAD LOAN CULPRIT STARTING TO LOOK LIKE?

Based on these very reasonable variables we start to see a story of a person of any income level, homeowner or not, who has low to no recent credit inquiries or delinquencies that has alot of debt and credit card debt that needs to be serviced better. Thats why they came to Lending Club to get help bridge the gap of their income versus the payments needed to stay afloat.

Upon a more detailed look based on the model that seems to explain the most of the bad loans, the culprit seems to be healthy, 5-20 open credit accounts, has a relatively low revolving balance, with loans that are 36 months.

Looking at the data and letting it guide me leads me to either say, either the profiles are actually healthy but face circumstancial reasons beyond their control that leads them

OR

Profile can be lying. either overreporting the good things like employment length, or under reporting bad things like credit inquiries or delinquencies

**IS THIS IT? Innocent unlucky people or liars are our profile?**

**I believe these are 2 good areas to start with and may be a mix of both.**

However upon actually building models much of the data competes with each other over explaining and not explaining things. They tended to pull the direction in different ways.

Each type of variable were significant in of themselves but overall didnt help much.

There were a few major ways I tried to attack the data

1. Throw everything into the blender and see what happens method (Hoping to get lucky!)

Result: Garbage, the resulting model did worse than assuming all loans were successful

1. Group up items that seemed to work together like income and employment length or inquiries and delinquecies

Result: Slightly better but very slight, predicted better by < 1%

3)(Let the data guide me method) Started over and threw away all preconceived notions of what may be significant and followed variables that were more significant than others according to the data.

Result: Humbling Experience. Variables that seemed important were not important

End result found a model that beat baseline by over 10%

*METHODOLOGY:*

I took all of the loan status that were fully paid off and set this as a new feature field called bad\_loans and set these at "0", then took all of the defaulted or charged off status loans as "bad loans" and set these to "1".

suppressWarnings(suppressMessages(library(ggplot2)))  
suppressWarnings(suppressMessages(library(dplyr)))  
suppressWarnings(suppressMessages(library(tidyr)))  
LC <- read.csv("LoanStats3d\_securev1.csv")  
  
 LC[c("bad\_loans")] = as.integer(0)  
  
 #Consolidate variables to speed up the data crunching  
   
 LC\_revised <- LC %>% select(loan\_amnt, term, int\_rate, installment, grade, sub\_grade, emp\_length, home\_ownership, annual\_inc, issue\_d, loan\_status, purpose, zip\_code, addr\_state, dti, delinq\_2yrs, fico\_range\_low, fico\_range\_high, inq\_last\_6mths, total\_pymnt, total\_rec\_prncp, total\_rec\_int, bad\_loans, open\_acc, pub\_rec, revol\_bal, revol\_util, total\_acc, collections\_12\_mths\_ex\_med, all\_util, tot\_hi\_cred\_lim)  
  
 LC\_filtered <- filter(LC\_revised, loan\_status == "Default" | loan\_status == "Fully Paid" | loan\_status == "Charged off")  
  
 default <- LC$loan\_status == "Default"  
 charged\_off <- LC$loan\_status == "Charged Off"  
   
 LC$bad\_loans[default] = as.integer(1)  
 LC$bad\_loans[charged\_off] = as.integer(1)

Took the subset data of only good loans and bad loans and split them into a training set and a testing set

library(caTools)  
set.seed(0)   
split = sample.split(LC\_filtered$bad\_loans, SplitRatio = 0.75)  
LC\_filtered\_Train <- subset(LC\_filtered, split == TRUE)  
LC\_filtered\_Test <- subset(LC\_filtered, split == FALSE)

Then Used a GLM Logistical model that predicts between 2 choices bad loans or a "1" flag or a good loans or a "0" flag. I used this as there were only 2 states I cared about, good loans and bad loans, which made this version ideal to use.

I didn't use a regression as there wasn't a range of points to plot and draw a best fitting line to get a model. It was a very binary choice

I could have potentially used a Kmeans methodology to try to extract profile types that can be grouped but given the dataset is already so rich , the lower hanging fruit is to try to see what value can the existing data can be used without trying to extract new features. However this can be a good next step to expand to once we have a base model to predict bad loans. So currently it is more pragmatic to use what we have first and try to minimize the manpower and resources to get a more accurate model which may or may not bring back ROI.

Here were a few I tried that seemed good:

#LC\_filtered\_log = glm(bad\_loans ~ delinq\_2yrs + inq\_last\_6mths, data = #LC\_filtered\_Train, family = binomial)   
#  
# AIC: 51985  
#  
#LC\_filtered\_log2 = glm(bad\_loans ~ emp\_length + delinq\_2yrs + inq\_last\_6mths, data = #LC\_filtered\_Train, family = binomial)  
#  
#LC\_filtered\_log3 = glm(bad\_loans ~ purpose + emp\_length + delinq\_2yrs + #inq\_last\_6mths, data = LC\_filtered\_Train, family = binomial)  
#  
#LC\_filtered\_log4 = glm(bad\_loans ~ addr\_state + purpose + emp\_length + delinq\_2yrs + #inq\_last\_6mths, data = LC\_filtered\_Train, family = binomial)  
#  
#LC\_filtered\_log5 = glm(bad\_loans ~ fico\_range\_low + addr\_state + purpose + emp\_length #+ delinq\_2yrs + inq\_last\_6mths, data = LC\_filtered\_Train, family = binomial)

I easily spent 40-80 hours looking at many models to find dead ends. Most of the models were dead ends. However with each iteration, slowly found some variables that played nice with each other!

Went through and tried to find variables that affected the AIC the most with significant variables. Also tried to make sure the direction of the data was in the same sign and direction and not competing against each other.

I went through several models to try to find the correct version

LC\_filtered\_log9

With the model that works that seemed to minimize the errors and explain the data the most, went to try to train and test the data to ensure its predictability.

So using the confusion Matrix to try to evaluate the accuracy of the training set vs what it actually was.

suppressWarnings(suppressMessages(library(ggplot2)))  
suppressWarnings(suppressMessages(library(dplyr)))  
suppressWarnings(suppressMessages(library(tidyr)))  
suppressWarnings(suppressMessages(library(lubridate)))  
suppressWarnings(suppressMessages(library(caTools)))  
suppressWarnings(suppressMessages(library(ROCR)))  
suppressWarnings(suppressMessages(library(pROC)))  
  
#predict\_train9 <- predict(LC\_filtered\_log9, type = "response")  
#table(LC\_filtered\_Train$bad\_loans, predict\_train9 > 0.5)  
   
# FALSE TRUE  
# 0 40238 1691  
# 1 2375 8023  
   
# (40238+8023)/(40238+1691+2375+8023)=.9222 accuracy  
# (40238+1691)/ (40238+1691+2375+8023)  
# [1] 0.8012881 (Baseline accuracy)  
  
# table(LC\_filtered\_Train$bad\_loans, predict\_train9 > 0.4)  
   
# FALSE TRUE  
# 0 39499 2430  
# 1 1804 8594  
   
# (39499+8594)/(39499+2430+1804+8594) =.919  
# table(LC\_filtered\_Train$bad\_loans, predict\_train9 > 0.3)  
   
# FALSE TRUE  
# 0 38525 3404  
# 1 1343 9055  
   
# (39525+9055)/(38525+3404+1343+9055)= .928 accuracy  
# table(LC\_filtered\_Train$bad\_loans, predict\_train9 > 0.2)  
   
# FALSE TRUE  
# 0 37045 4884  
# 1 909 9489  
   
# (37045+9489)/(37045+4884+909+9489)- .889 accuracy  
   
# table(LC\_filtered\_Train$bad\_loans, predict\_train9 > 0.1)  
   
# FALSE TRUE  
# 0 34020 7909  
# 1 526 9872  
   
# (34020+9872)/(34020+7909+526+9872) = .838 accurracy  
# (34020+7909)/(34020+7909+526+9872) = .80 dominant bias

Keep in mind the dominant bias of the data is about 80%~ good loans so our accuracy needs to be more accurate than 80% to beat the baseline of assume all loans are good and getting 80% of the loans correct.

We have a gain improvement of little over 11% but is the threshold of 0.5 sufficient?

*WAY TO PICK A THRESHOLD*

There are multiple ways to look at it but goal is to make the most money or predict as many bad loans so Lending Club can charge fees to borrowers and the Lenders can feel safer knowing the money they invest to loan will earn a return with minimal losses.

That is if the goal is to increase trust is to eliminate as many potential bad loans at the expense of good loans.

However once confidence is back, can allow the standards to be more lax to grow revenue for LC.

# good\_loans\_data <- filter(LC\_filtered, loan\_status == "Fully Paid")  
# bad\_loans\_data <- filter(LC\_filtered, bad\_loans == "1")  
# nrow(bad\_loans\_data)  
#[1] 13864  
#   
# good\_loans\_data <- filter(LC\_filtered, bad\_loans == "0")  
# nrow(good\_loans\_data)  
#[1] 55905  
#   
# average\_good\_loan\_prof <- (sum(good\_loans\_data$total\_pymnt) - sum(good\_loans\_data$loan\_amnt))/nrow(good\_loans\_data)  
#   
# average\_good\_loan\_prof  
#[1] 1100.512  
#   
# average\_bad\_loan\_prof <- (sum(bad\_loans\_data$total\_pymnt)- sum(bad\_loans\_data$loan\_amnt))/nrow(bad\_loans\_data)  
#   
# average\_bad\_loan\_prof  
#[1] -11907.74  
#  
# -11907.74/1100.512  
#[1] -10.82018

so seems on an aggregate bad loans can impact your profitability by almost 11x  
i would argue you need to make sure need to pick 12 good loans for every 1 bad to break even on principal. Thats not counting the opportunity cost lost and any time/efforts lost

I would suggest a between .2 to .3 threshold range to maximize the # of bad loans found without taking out too many productive loans by flagging them bad incorrectly.

#predict\_test9 <- predict(LC\_filtered\_log9, type = "response", newdata = LC\_filtered\_Test)  
#  
#table(LC\_filtered\_Test$bad\_loans, predict\_test9 > 0.4)  
#   
# FALSE TRUE  
# 0 13195 781  
# 1 561 2905  
#  
# (2905+13195)/(13195+781+561+2905) = .923 accuracy  
#  
#table(LC\_filtered\_Test$bad\_loans, predict\_test9 > 0.3)  
# FALSE TRUE  
# 0 12866 1110  
# 1 420 3046  
#   
# (3046+12866)/(12866+1110+420+3046) = .912 accuracy  
# (12866+1110)/(12866+1110+420+3046) = .80 dominant bias  
#   
#table(LC\_filtered\_Test$bad\_loans, predict\_test9 > 0.2)  
#   
# FALSE TRUE  
# 0 12353 1623  
# 1 295 3171  
# (12353+3171)/(12353+1623+295+3171) = .890 accuracy

So in the end a 0.3 Threshold was chosen and ran against the test set to ensure the same model predicts with new data and it does!

11% better than the baseline

*EVALUATION/IMPLICATIONS*

Originally I was looking for people that looked unhealthy that eventually defaulted, looking for credit inquiries, delinquencies, low income, low employment lengths and long term loans with high interest rates

Instead it seems the type that defaults may be employed for a long time, fairly healthy, relatively low revolving balances, relatively low open accounts with low to no credit inquires.

Why may this be?

It seems very much the opposite of the type that may default as if you're looking for a person that may commit a crime, you may check for tattoes, previous jail time or arrests.

Seems counterintuitive especially when I looked at it.

Here are some ideas that may play a role:

The people who default or have a charge off were intending to pay back in full and have the background that says they have no problem to do so. But why default if that is the case?

1. External circumstances out of their control like getting sick suddenly.
2. Hiding the fact they are not okay by under-reporting or the reporting data has a delay
3. Complacency as it seems obvious the debt/credit card debt will get paid eventually with the new Lending Club consolidated Loan

*SUGGESTIONS/RECOMMENDATIONS FOR LENDING CLUB:*

1. since total payments seems to explain why loans go bad, may check for a longer history of on time payments to judge a "safe" grade A or B loan or give a higher risk without verifiable on time payment histories.

2)There seems to be a sense of complacency based on the data that may make what seems to be an a low risk profile be a hidden higher risk. I'd try to offer small discounts to accelerate payments who pay early instead of last minute.

3)Other miscellaneous External Circumstancial risks. This may be the product of the current low interest environment or related Macroeconomic factors. There should be a survey done to find out why the current ones to find out the circumstances on how the charge off happens, whether it was controllable risk or not.

1. Motives on liars would be a good study to make long term

Overall I feel the model to identify bad loans should biased to eliminating bad loans in the short term to win back confidence of the public as tweaks are made to close the hidden risks that were graded too generously safe.

APPENDIX

--Here were somee approaches that were tested but didnt seem to have desired effects

--Kmeans to strenghen the signal for Purpose

--Looking at the financial data that many typically may think will work

--Using too many variables

Variables I thought were important:

*income* employment length *Purpose of the loan* Delinquencies \*Inquiries

Variables that are actually important:

*Total Payments* Accounts Open *Grade* collections on medical \*revolving Balance

Based on the variables that were actually important in predicting bad loans, it seems the variables that were important were variables that hint of how consistency seems to be the main reason why people pay it off successfully. For the ones that don't seem to hint of things that were beyond their control like collections medical debt or # of accounts or revolving balance that hint some external factor overwhelmed them.

WAYS TO IMPROVE THE MODEL:

Seems once you know payments made or total payments made, next most reasonable way to improve the model is to create a linear regression based on payments to improve this aspect.