Data Science for Business
Lecture #6
Decision Tree Examples

Prof. Alan L. Montgomery

The University of Hong Kong & Carnegie Mellon University, Tepper School of Business

email: alanmontgomery@cmu.edu

Lending Club In-Class Exercise

Review from last session with Logistic Regression

LendingClub

Sign in Help

Personal Loans Auto Refinancing Business Loans Patient Solutions Investing How It Works About Us

How does an online credit marketplace work?

Lending Club uses technology to operate a credit marketplace at a lower cost than traditional bank loan programs, passing the savings on to borrowers in the form of lower rates and to investors in the form of solid returns. Borrowers who used a personal loan via Lending Club to consolidate debt or pay off high interest credit cards report in a survey that the interest rate on their loan was an average of 25% lower than they were paying on their outstanding debt or credit cards.[1]

By providing borrowers with better rates, and investors with attractive, risk-adjusted returns, Lending Club has earned among the highest satisfaction ratings in the financial services industry.[2]

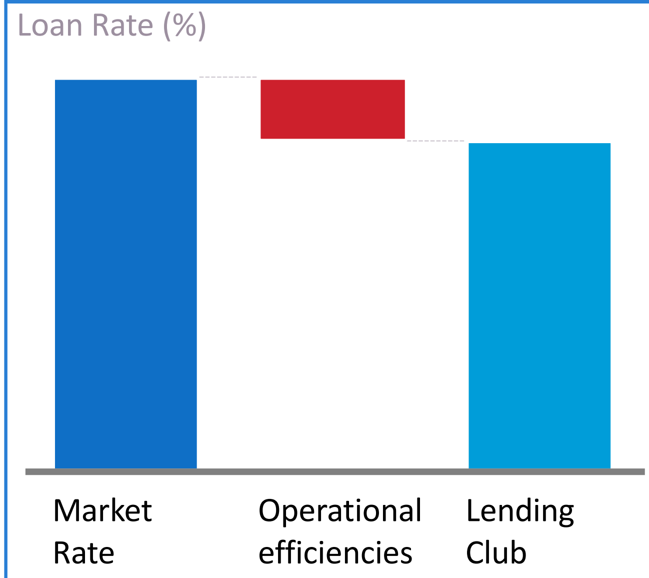
Introduction to Lending Club

Fast facts

- Lending Club is a peer-to-peer lending system that has set up an online marketplace connecting investors to borrowers
- Lending Club operates at a lower cost than traditional bank lending programs and pass the savings on to borrowers (lower rates) and to investors (solid returns)
- In 2007, Lending Club made 9,758 loans with ~$75M in loan value
- All loan lengths are 3 years
- Investors shared in $15M in profits after accounting for $12.5M in loan default losses

CEO has ask you to determine if there is a better model for determining credit worthiness

Average loan rate offered



Two-sided Market

Better for Borrowers

We cut the cost and complexities of traditional bank loans and pass the savings on to borrowers.

Learn more

- **Easy online application**
- **Low fixed rates**
- **Fixed monthly payments**
- **Flexible terms**
- **No prepayment penalties**
- **No hidden fees**
- **Friendly service**

Better for Investors

At Lending Club you can earn attractive risk-adjusted returns by quickly and easily investing in a diversified portfolio of loans.

Learn more

- **Solid returns with historical returns by Grade A-C of 5.01% to 7.38%.[5]**
- **Monthly cash flow**
- **Simple and straightforward**
- **Easy to diversify across many Loans**
- **401(k) rollover and retirement accounts available**

25%

Borrowers reduce their rates by an average of 25%[1]!

Borrowers who used a personal loan[*] via Lending Club to consolidate debt or pay off high interest credit cards report in a survey that the interest rate on their loan was an average of 25% lower than they were paying on their outstanding debt or credit cards.

Solid Returns

5-7%

Lending Club Notes have historical annual returns between 5% and 7%. Each Note represents a fraction of an underlying loan.[1]

Low Volatility

99%

99% of investors who invest in 100+ Notes of relatively equal size have seen positive returns.[2]

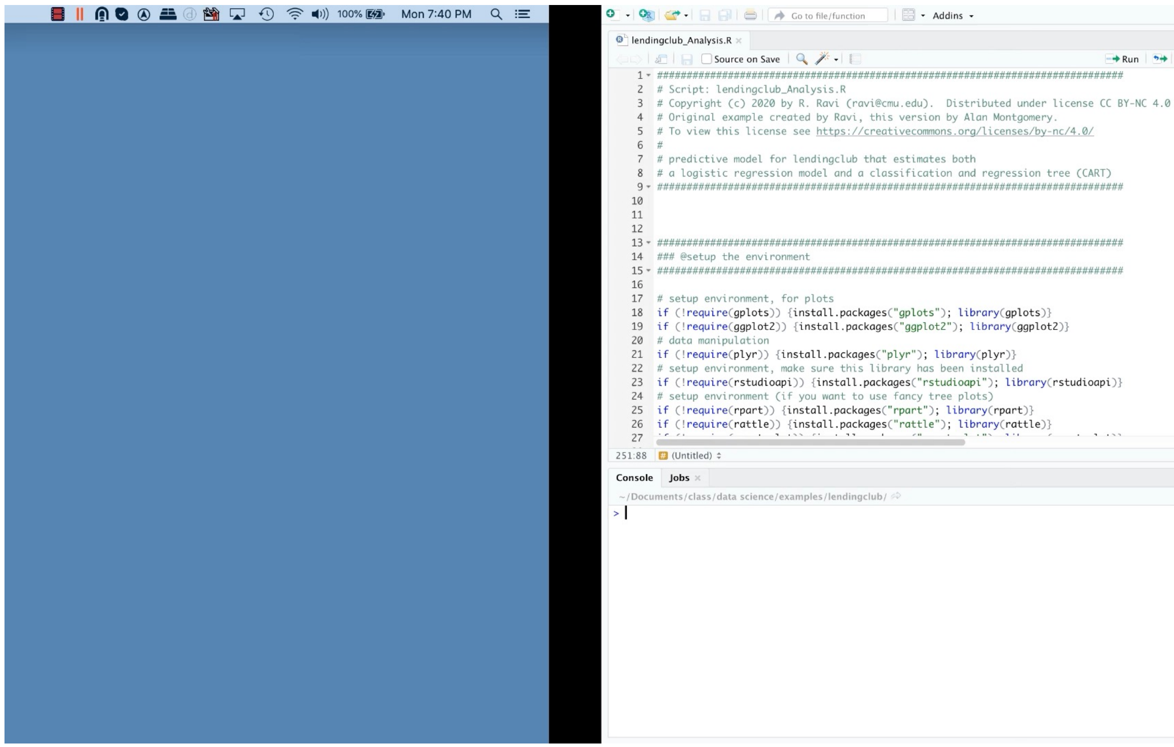
Monthly Cash Flow

2-5%

Investors receive between 2-5% of their total investment back in cash payments as borrowers make their monthly loan payment.[3]

Lending Club Dataset

Variable	Description
default	1 if the customer did not fully pay back the loan, and 0 otherwise.
credit.policy	1 if the customer meets the credit underwriting criteria of LendingClub.com, and 0 otherwise.
purpose	The purpose of the loan (takes values "credit_card", "debt_consolidation", "educational", "major_purchase", "small_business", and "all_other").
int.rate	The interest rate of the loan, as a proportion (a rate of 11% would be stored as 0.11). Borrowers judged by LendingClub.com to be more risky are assigned higher interest rates.
installment	The monthly installments ($) owed by the borrower if the loan is funded.
log.annual.inc	The natural log of the self-reported annual income of the borrower.
dti	The debt-to-income ratio of the borrower (amount of debt divided by annual income).
fico	The FICO credit score of the borrower.
days.with.cr.line	The number of days the borrower has had a credit line.
revol.bal	The borrower's revolving balance (amount unpaid at the end of the credit card billing cycle).
revol.util	The borrower's revolving line utilization rate (the amount of the credit line used relative to total credit available).
inq.last.6mths	The borrower's number of inquiries by creditors in the last 6 months.
delinq.2yrs	The number of times the borrower had been 30+ days past due on a payment in the past 2 years.
pub.rec	The borrower's number of derogatory public records (bankruptcy filings, tax liens, or judgments).



100% Mon 7:40 PM
Go to file/function
Addins
lendingclub_Analysis.R
Source on Save
Run
##
Script: lendingclub_Analysis.R
Copyright (c) 2020 by R. Ravi (ravi@cmu.edu). Distributed under license CC BY-NC 4.0
Original example created by Ravi, this version by Alan Montgomery.
To view this license see https://creativecommons.org/licenses/by-nc/4.0/
#
predictive model for lendingclub that estimates both
a logistic regression model and a classification and regression tree (CART)
##
##
@setup the environment
##
# setup environment, for plots
if (!require(gplots)) {install.packages("gplots"); library(gplots)}
if (!require(ggplot2)) {install.packages("ggplot2"); library(ggplot2)}
# data manipulation
if (!require(plyr)) {install.packages("plyr"); library(plyr)}
# setup environment, make sure this library has been installed
if (!require(rstudioapi)) {install.packages("rstudioapi"); library(rstudioapi)}
# setup environment (if you want to use fancy tree plots)
if (!require(rpart)) {install.packages("rpart"); library(rpart)}
if (!require(rattle)) {install.packages("rattle"); library(rattle)}
251:88
(Untitled)
Console
Jobs
~/Documents/class/data science/examples/lendingclub/
>

Explaining the Model
The most important reasons for default…

Variable	Estimate	Importance	Why consumers *more* likely to default if they…
installment	▲ 0.00	0.28	have high payments
log.annual.inc	▼ -0.48	0.26	have low incomes
inq.last.6mths	▲ 0.10	0.25	have many recent inquiries to borrow
purposecredit_card	▼ -0.74	0.22	are not refinancing credit cards
fico	▼ -0.01	0.19	have poor credit scores
purposedebt_consolidation	▼ -0.42	0.18	are not refinancing other debt
int.rate	▲ 4.91	0.14	have high interest rates
purposesmall_business	▲ 0.48	0.13	are refinancing for small business loans
revol.util	▲ 0.00	0.13	are using higher percentage of available revolving credit
credit.policy	▼ -0.32	0.12	do not meet current credit policy
revol.bal	▲ 0.00	0.10	have high revolving credit balances
pub.rec	▲ 0.33	0.09	have previous bankruptcy or default
purposeeducational	▼ -0.23	0.04	are not refinancing educational loans
purposemajor_purchase	▼ -0.14	0.03	are not refinancing major purchases
purposehome_improvement	▲ 0.01	0.00	are refinancing home important loans

Top 5 reasons that contribute to default

Lending Club

Predicting Default using Decision Trees

In-Class Exercise: Part 3
Decision Tree

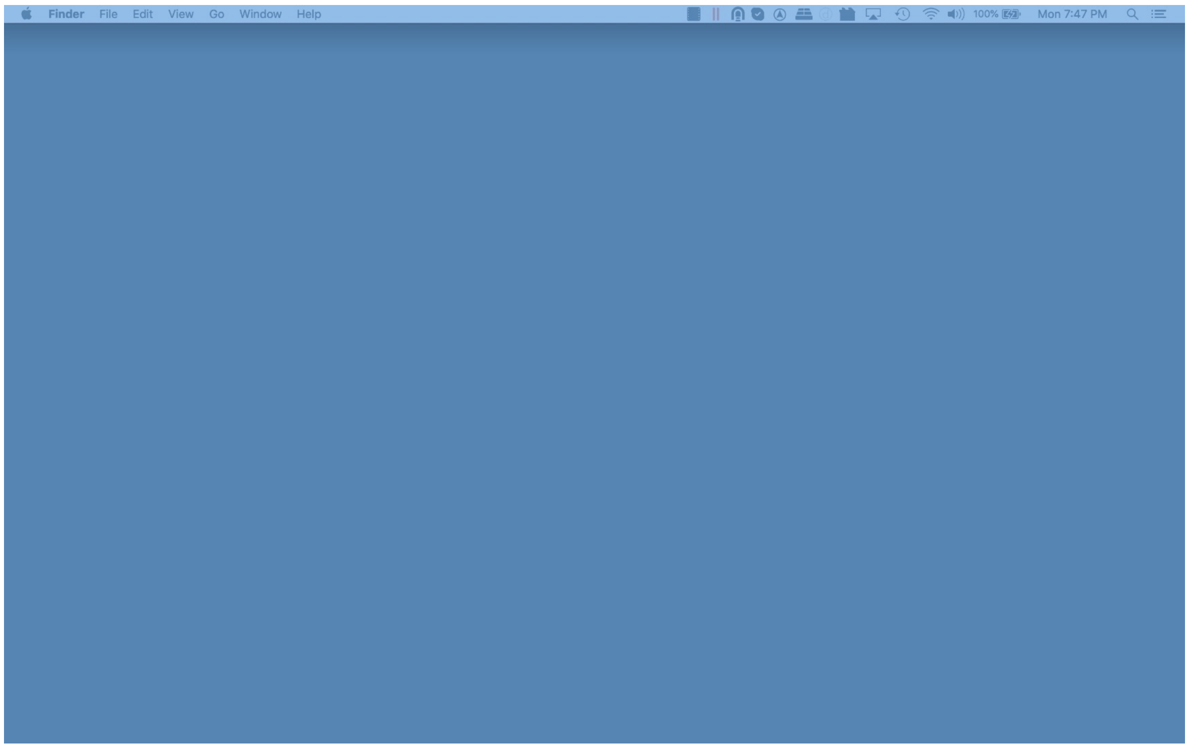
Step through the "@tree" in the script and carefully consider the results.

What is a good tree? (e.g., What value for cp?)

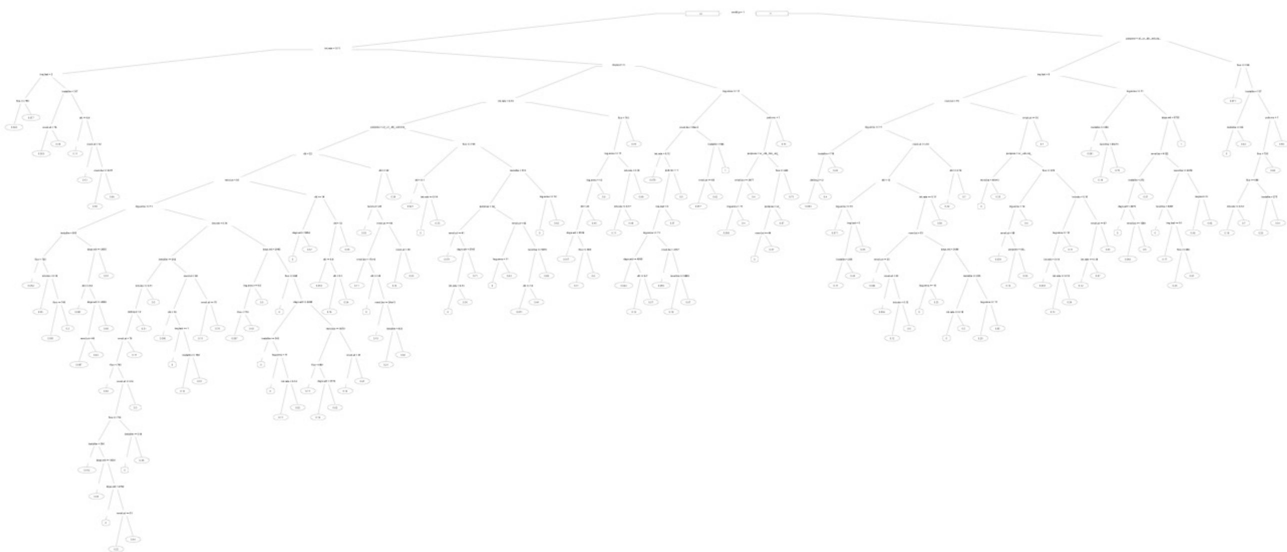
Review your initial hypothesis? What did you learn from the decision tree about default?

Complete the following two slides:

- Explaining your model
- Use your model to construct three different classification matrices


Finder
File
Edit
View
Go
Window
Help
100%
Mon 7:47 PM

Decision Tree: Estimate Full Tree cp=.001
Do we need to prune this tree back?



Decision Tree: Estimate Full Tree cp=.001
Compare training and validation samples

Training Sample

```
$confmatrix
         trueclass
predclass    0    1
        0 4048  251
        1  757  641

$accuracy
[1] 0.8230648

$truepos
[1] 0.7186099

$precision
[1] 0.4585122

$trueneg
[1] 0.8424558
```

Notice the large drop in performance. This is a clear indicator of over-fitting.

Test Sample

```
$confmatrix
         trueclass
predclass    0    1
        0 2514  429
        1  726  212

$accuracy
[1] 0.7023963

$truepos
[1] 0.3307332

$precision
[1] 0.2260128

$trueneg
[1] 0.7759259
```

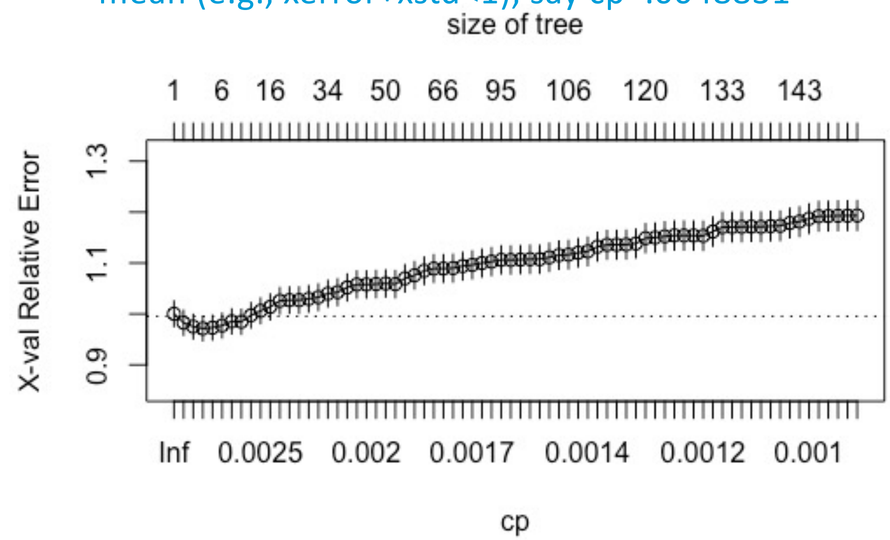
Decision Tree: Prune the tree to cp=.0048831
This is only a rule of thumb, could try many settings

```
printcp(ctree.full)
plotcp(ctree.full)
```

	CP	nsplit	rel error	xerror	xstd
1	0.0252911	0	1.00000	1.00036	0.025050
2	0.0113258	1	0.97471	0.98267	0.024495
3	0.0061745	2	0.96338	0.97510	0.024068
4	0.0048831	3	0.95721	0.97144	0.023974
5	0.0046170	4	0.95233	0.97303	0.023989
6	0.0032407	5	0.94771	0.97729	0.024103
7	0.0030520	8	0.93799	0.98544	0.024394
8	0.0030470	11	0.92883	0.98461	0.024354
9	0.0025003	12	0.92578	0.99736	0.024745
10	0.0024418	13	0.92328	1.00615	0.025109
11	0.0023533	15	0.91840	1.01386	0.025326
12	0.0022901	19	0.90899	1.02583	0.025661
13	0.0022776	23	0.89970	1.02721	0.025689
14	0.0022395	24	0.89742	1.02721	0.025689

cp=complexity
rel error=error relative to base error (one leaf)
xerror=relative error for cross-validated model
xstd=measure of uncertainty for mean

A good choice of cp is to choose the smallest model that is within a one standard error of the mean (e.g., xerror+xstd<1), say cp=.0048831



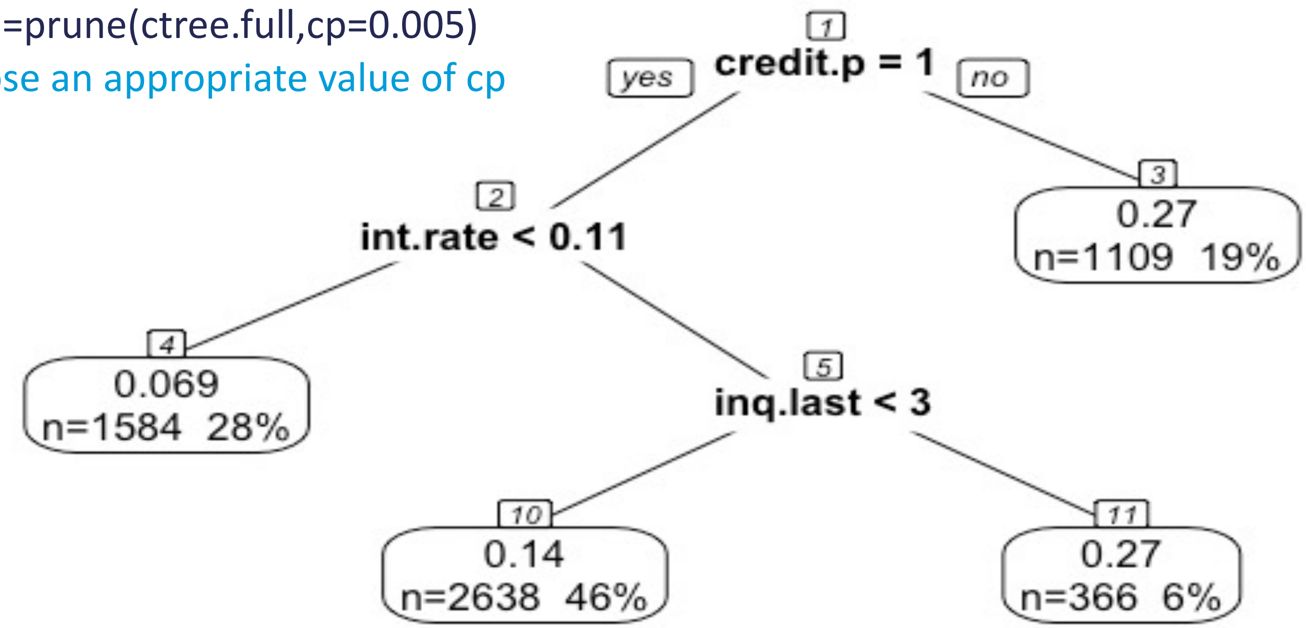
Another good choice of cp for pruning is the one with the minimum “xerror”, say cp=.0032407

Decision Tree: Tree with cp=.005
Easy to understand tree

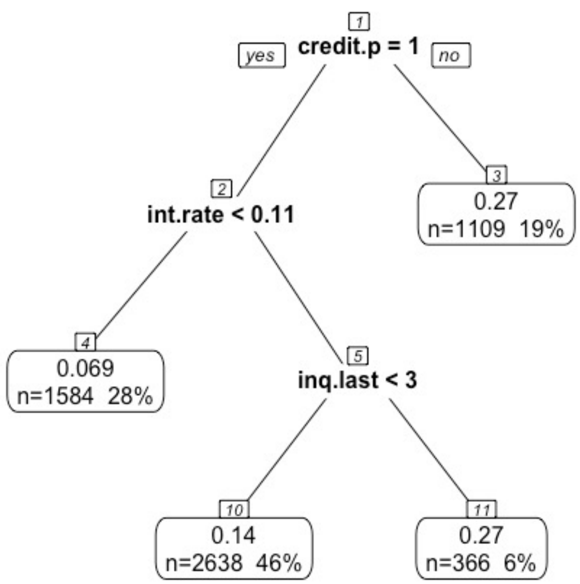
Prune the tree using the "prune" command:

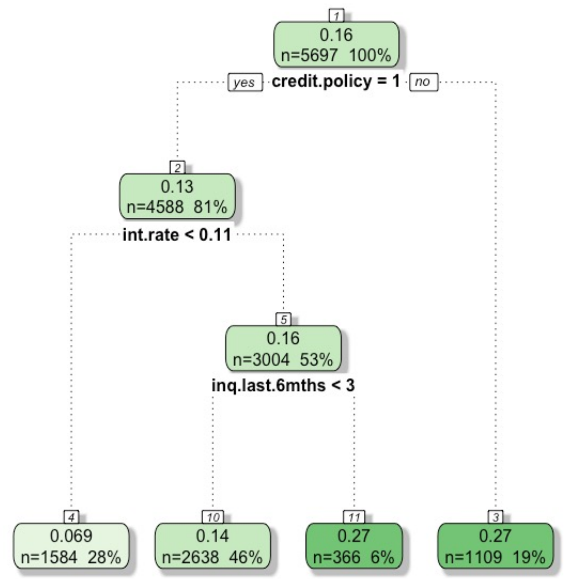
ctree=prune(ctree.full,cp=0.005)

Choose an appropriate value of cp



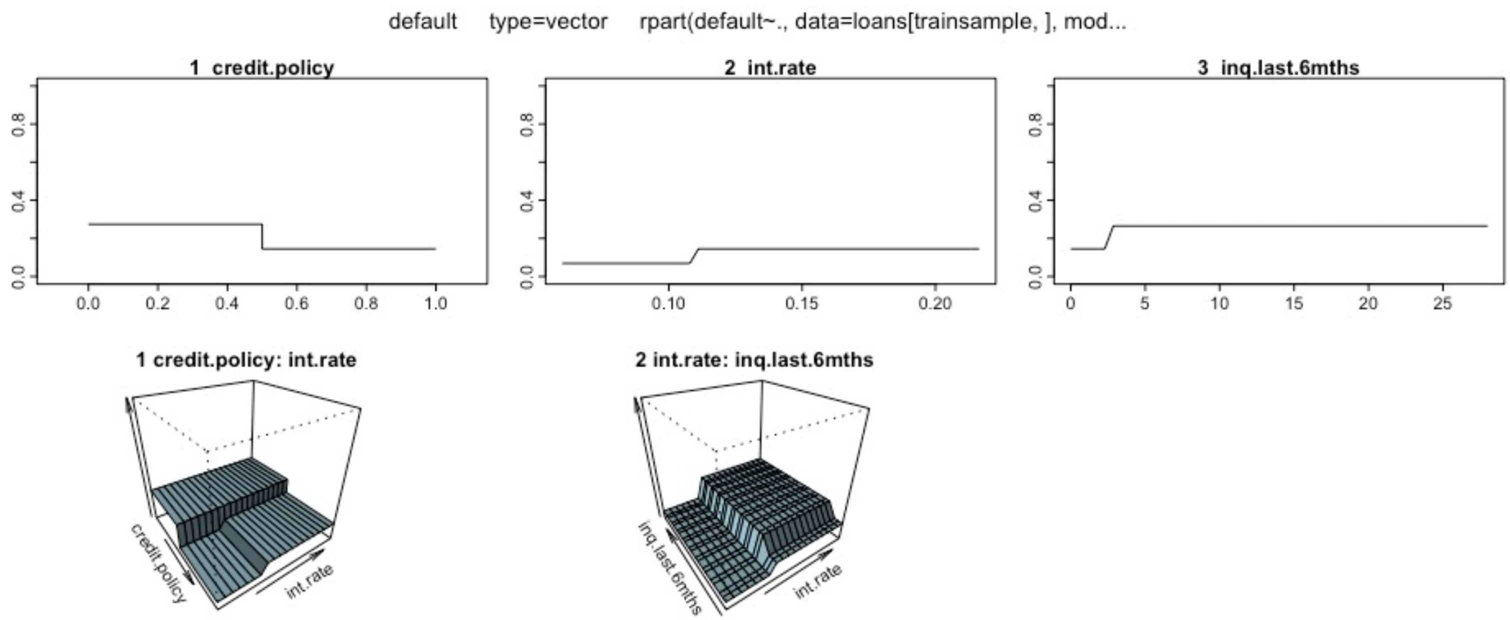
Explaining the Model





Use this slide to explain your tree model: you can show the tree, call out the most important predictors, or otherwise justify the rules highlighted by the tree

Decision Tree
Use plotmo *to understand response curve*



Examine Prediction Accuracy using Confusion Matrix

Build a confusion matrix for cutoff of 0.16 (above or below average)
Make sure that the confusion matrix is for the test set and not the training set for the model

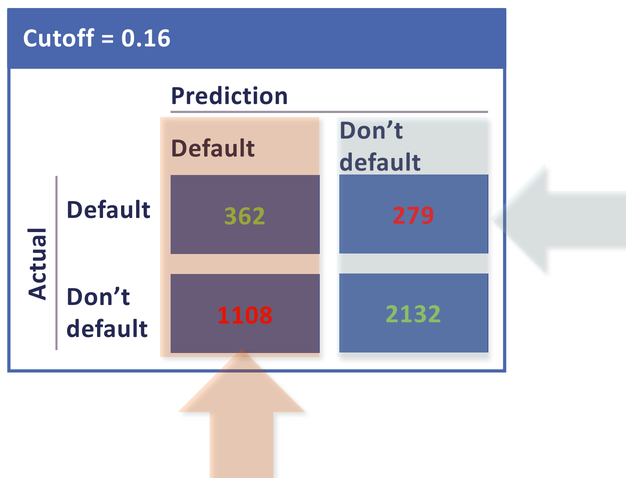
Cutoff = 0.16

		Prediction	
		Default	Don’t default
Actual	Default		
	Don’t default		

Examine Prediction Accuracy using Confusion Matrix

Build a confusion matrix for cutoff of 0.16 (above or below average)
Make sure that the confusion matrix is for the test set and not the training set for the model

```
$confmatrix
         trueclass
predclass    0    1
        0 2132  279
        1 1108  362

$accuracy
[1] 0.6426179

$truepos
[1] 0.5647426

$precision
[1] 0.2462585

$trueneg
[1] 0.6580247

$lift
[1] 1.976695
```



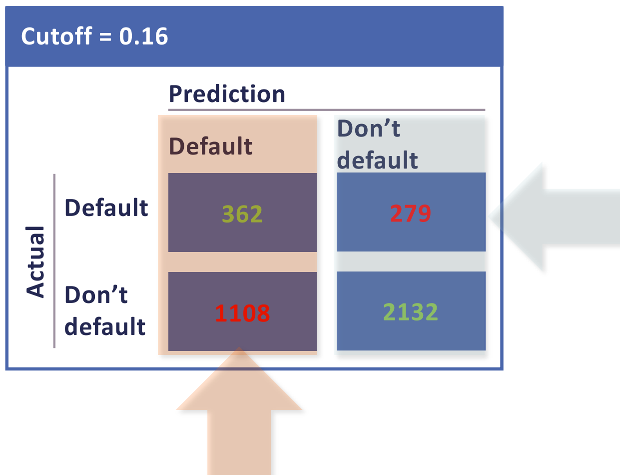
If we predict "don't default" then we will make loans, so profits come from those that don't default (279) and losses come from those that default (2132)

If we predict "default" then we will not make loans, so no profits

Examine Prediction Accuracy using Confusion Matrix

Build a confusion matrix for cutoff of 0.16 (above or below average)
Make sure that the confusion matrix is for the test set and not the training set for the model

```
$confmatrix
         trueclass
predclass    0    1
        0 2132  279
        1 1108  362

$accuracy
[1] 0.6426179

$truepos
[1] 0.5647426

$precision
[1] 0.2462585

$trueneg
[1] 0.6580247

$lift
[1] 1.976695
```



If we predict "don't default" then we will make loans, so profits come from those that don't default (279) and losses come from those that default (2132)

If we predict "default" then we will not make loans, so no profits

Decision Tree: Compare performance
Which tree is better?

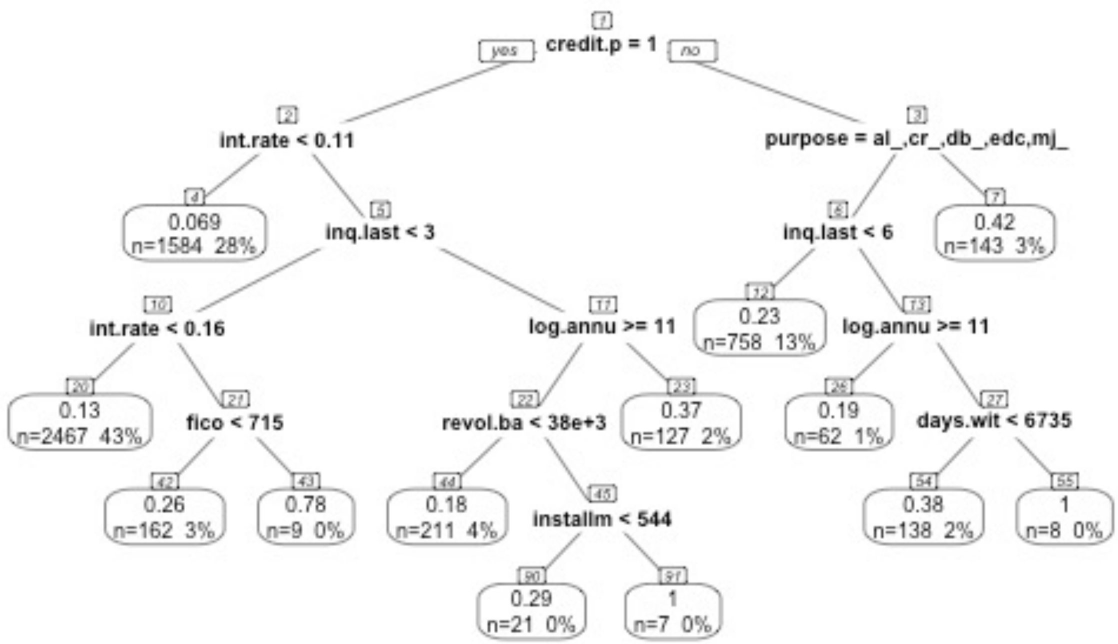
cp=.005

```
$confmatrix
          trueclass
predclass    0    1
        0 2514  429
        1  726  212

$accuracy
[1] 0.7023963

$truepos
[1] 0.3307332

$precision
[1] 0.2260128

$trueneg
[1] 0.7759259
```

cp=.001

```
$confmatrix
          trueclass
predclass    0    1
        0 2489  382
        1  751  259

$accuracy
[1] 0.7080649

$truepos
[1] 0.4040562

$precision
[1] 0.2564356

$trueneg
[1] 0.7682099
```

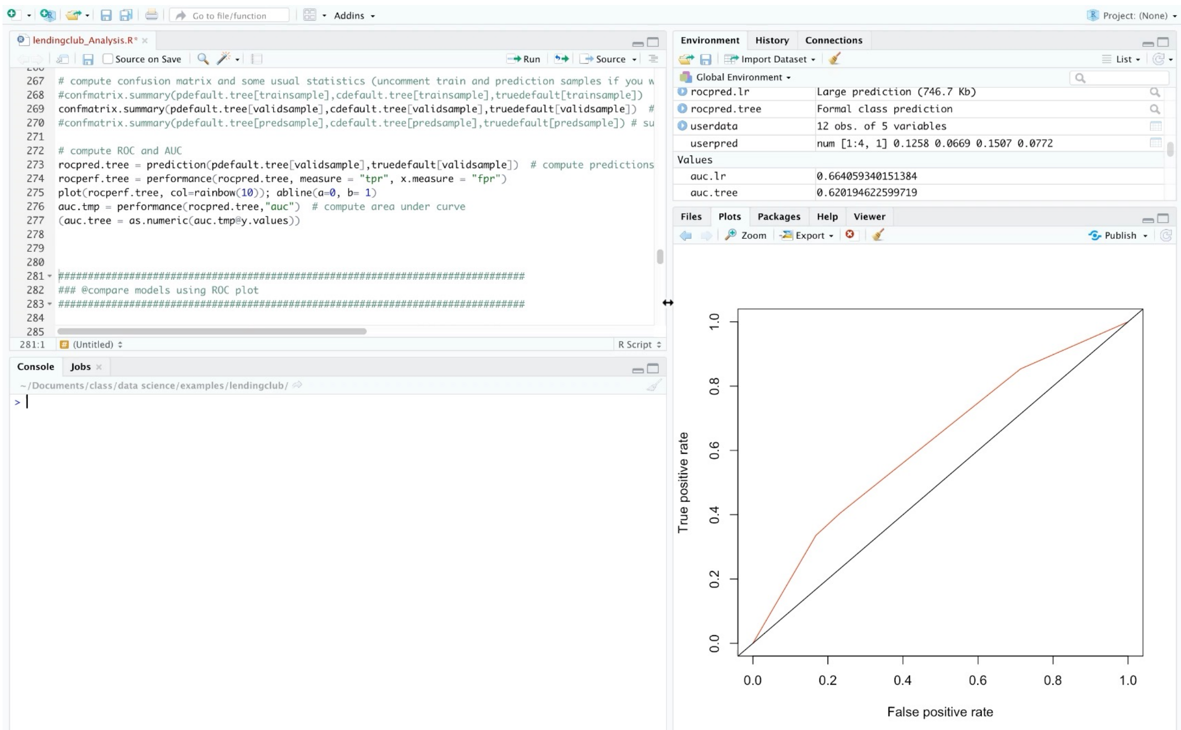
Decision Tree
Alternative tree with cp=0.0030470




```
$confmatrix
         trueclass
predclass    0    1
        0 2391  355
        1  849  286

$accuracy
[1] 0.6897707

$truepos
[1] 0.4461778

$precision
[1] 0.2519824

$trueneg
[1] 0.737963
```

Lending Club

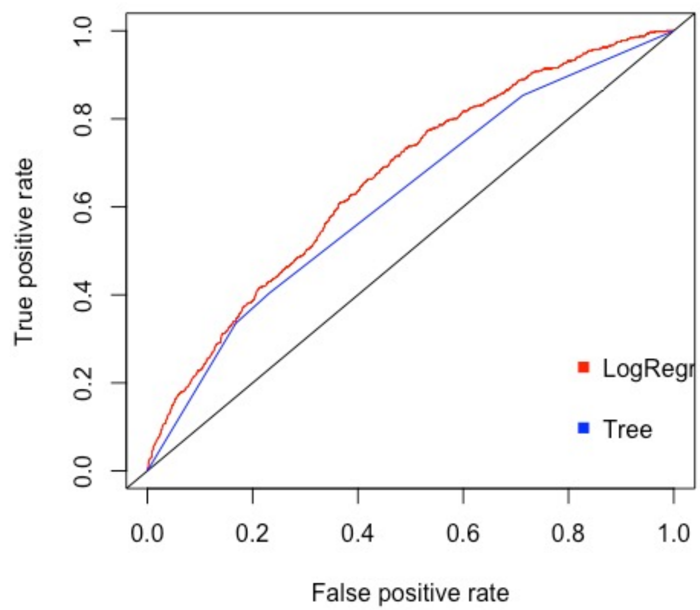
Which model to use?


lendingclub_Analysis.R*
Source on Save
Run
Source
compute confusion matrix and some usual statistics (uncomment train and prediction samples if you w
#confmatrix.summary(pdefault.tree[trainsample],cdefault.tree[trainsample],truedefault[trainsample])
confmatrix.summary(pdefault.tree[validsample],cdefault.tree[validsample],truedefault[validsample]) #
#confmatrix.summary(pdefault.tree[predsample],cdefault.tree[predsample],truedefault[predsample]) # su
# compute ROC and AUC
rocpred.tree = prediction(pdefault.tree[validsample],truedefault[validsample]) # compute predictions
rocperf.tree = performance(rocpred.tree, measure = "tpr", x.measure = "fpr")
plot(rocperf.tree, col=rainbow(10)); abline(a=0, b= 1)
auc.tmp = performance(rocpred.tree,"auc") # compute area under curve
(auc.tree = as.numeric(auc.tmp@y.values))
###
@compare models using ROC plot
###
Console
Jobs
~/Documents/class/data science/examples/lendingclub/
Environment
History
Connections
Import Dataset
Global Environment
rocpred.lr Large prediction (746.7 Kb)
rocpred.tree Formal class prediction
userdata 12 obs. of 5 variables
userpred num [1:4, 1] 0.1258 0.0669 0.1507 0.0772
Values
auc.lr 0.664059340151384
auc.tree 0.620194622599719
Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

In-Class Exercise: Part 4
Choose a model

Compare the model on the various metrics shown in the following page and ROC curves generated by the "@compare" portion of the script.

Which model should we use?



Compare the models

Variable	Logistic Regression	Decision Tree
Accuracy	.643	.708
Precision	.246	.256
Recall	.565	.404
Lift in top decile	1.977	1.715
AUC	.664	.620

Lending Club

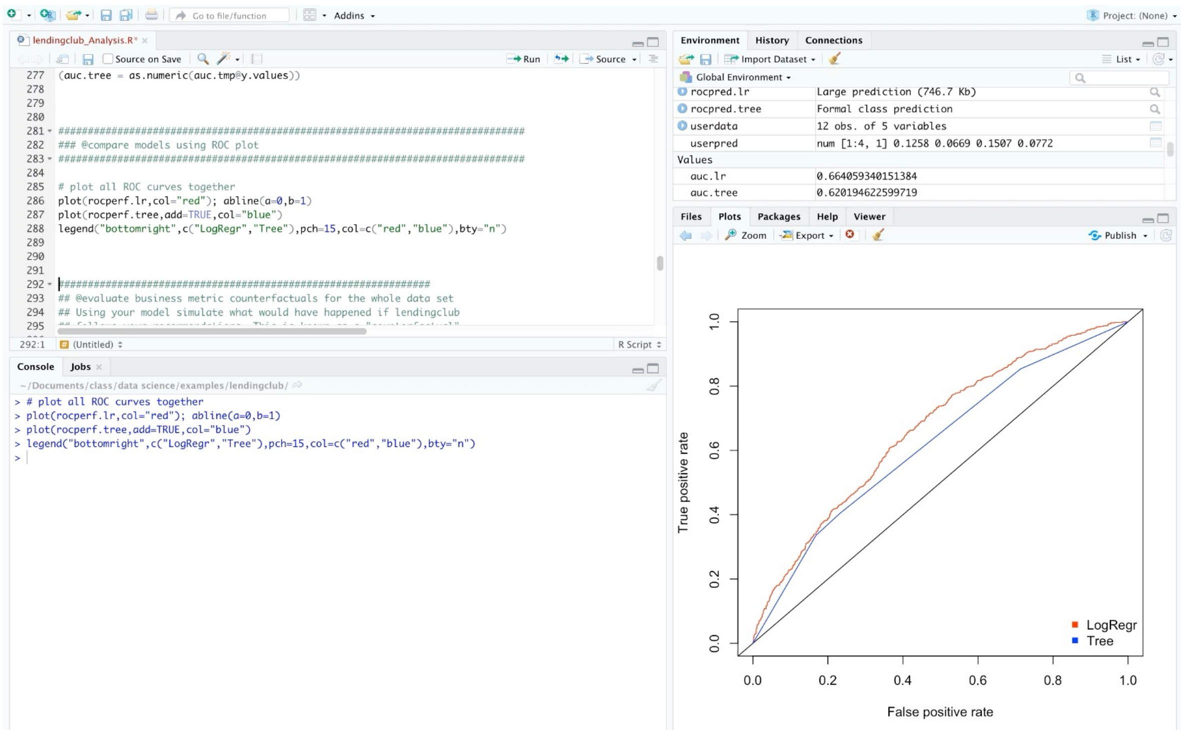
What cutoff to use?

In-Class Exercise: Part 5
Use the model to recommend loans

How can use use your model to improve the recommendation of which customers receive loans?

What KPIs or metrics should we use to judge the model?

If we follow your model recommendation then how much can we make?


lendingclub_Analysis.R*
Source on Save
Run
Source
(auc.tree = as.numeric(auc.tmp@y.values))
##
@compare models using ROC plot
##
# plot all ROC curves together
plot(rocperf.lr,col="red"); abline(a=0,b=1)
plot(rocperf.tree,add=TRUE,col="blue")
legend("bottomright",c("LogRegr","Tree"),pch=15,col=c("red","blue"),bty="n")
##
@evaluate business metric counterfactuals for the whole data set
Using your model simulate what would have happened if lendingclub
Console
Jobs
~/Documents/class/data science/examples/lendingclub/
> # plot all ROC curves together
> plot(rocperf.lr,col="red"); abline(a=0,b=1)
> plot(rocperf.tree,add=TRUE,col="blue")
> legend("bottomright",c("LogRegr","Tree"),pch=15,col=c("red","blue"),bty="n")
Environment
History
Connections
Import Dataset
Global Environment
rocpred.lr Large prediction (746.7 Kb)
rocpred.tree Formal class prediction
userdata 12 obs. of 5 variables
userpred num [1:4, 1] 0.1258 0.0669 0.1507 0.0772
Values
auc.lr 0.664059340151384
auc.tree 0.620194622599719
Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
Project: (None)
Addins
True positive rate
False positive rate
LogRegr
Tree
0.0
0.2
0.4
0.6
0.8
1.0

Examine Prediction Accuracy (by varying cutoff threshold for prediction)

Build a confusion matrix for 3 different cutoff thresholds
Make sure that the confusion matrix is for the test set and not the training set for the model

Cutoff = Lower

		Prediction	
		Default	Don't default
Actual	Default		
	Don't default		

Cutoff = 0.16

		Prediction	
		Default	Don't default
Actual	Default		
	Don't default		

Cutoff = Higher

		Prediction	
		Default	Don't default
Actual	Default		
	Don't default		

Examine Prediction Accuracy (by varying cutoff threshold for prediction)

Build a confusion matrix for 3 different cutoff thresholds
Make sure that the confusion matrix is for the test set and not the training set for the model

Cutoff = Lower = 0.10

Actual \ Prediction	Default	Don't default
Default	550	91
Don't default	2192	1048

Accuracy = 41%
TPR = 86%
Precision = 20%

Cutoff = 0.16

Actual \ Prediction	Default	Don't default
Default	362	279
Don't default	1108	2132

Accuracy = 64%
TPR = 56%
Precision = 25%

Cutoff = Higher = 0.25

Actual \ Prediction	Default	Don't default
Default	160	481
Don't default	363	2877

Accuracy = 78%
TPR = 25%
Precision = 31%

Connection between Profits and Confusion Matrix

Cutoff = 0.16

Actual \ Prediction	Default	Don't default
Default	362	279
Don't default	1108	2132

$$
\begin{aligned}
&E[Profit]= \\
&(Interest_{dd})\times \Pr\left(Actual{=}Don't\ Default \mid Predict{=}Don't\ Default\right)\times N \\
&-(Principal_{d})\times \Pr\left(Actual{=}Default \mid Predict{=}Don't\ Default\right)\times N \\
&=(Interest_{dd})\times\left(\frac{2132}{279+2132}\right)\times N-(Principal_{d})\times\left(\frac{279}{279+2132}\right)\times N \\
&=(Interest_{dd})\times 2132-(Principal_{d})\times 279
\end{aligned}
$$

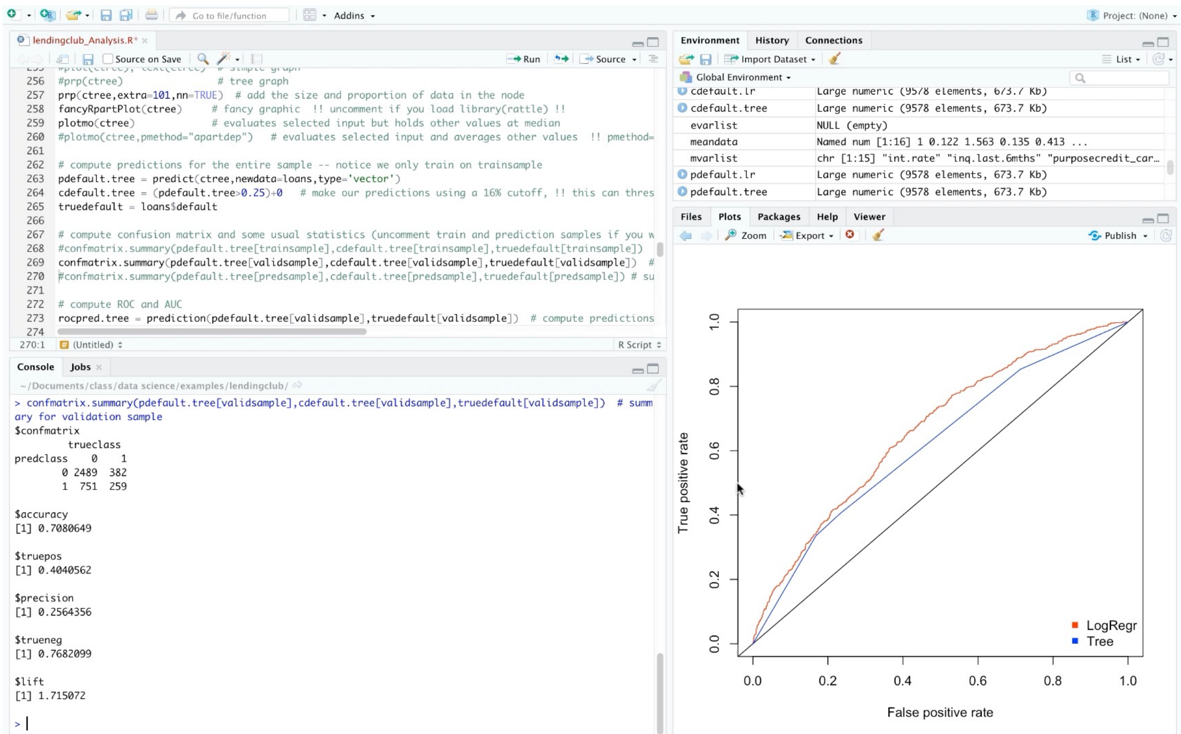
$$\text{where } N = 279+2132 = 2411$$

Another metric that we might be interested in Return on Invested Capital (ROIC):

$$ROIC = \frac{Interest_{dd}}{Principal_{d} + Principal_{dd}}$$

Profitability for Lending Club

What is the effect of different cutoffs on profits?


Go to file/function
Addins
Project: (None)
lendingclub_Analysis.R*
Source on Save
Run
Source
#prp(ctree)   # tree graph
prp(ctree,extra=101,nn=TRUE) # add the size and proportion of data in the node
fancyRpartPlot(ctree)   # fancy graphic !! uncomment if you load library(rattle) !!
plotmo(ctree)   # evaluates selected input but holds other values at median
#plotmo(ctree,pmethod="apartdep")   # evaluates selected input and averages other values !! pmethod=
# compute predictions for the entire sample -- notice we only train on trainsample
pdefault.tree = predict(ctree,newdata=loans,type='vector')
cdefault.tree = (pdefault.tree>0.25)+0   # make our predictions using a 16% cutoff, !! this can thres
truedefault = loans$default
# compute confusion matrix and some usual statistics (uncomment train and prediction samples if you w
#confmatrix.summary(pdefault.tree[trainsample],cdefault.tree[trainsample],truedefault[trainsample])
confmatrix.summary(pdefault.tree[validsample],cdefault.tree[validsample],truedefault[validsample]) #
#confmatrix.summary(pdefault.tree[predsample],cdefault.tree[predsample],truedefault[predsample]) # su
# compute ROC and AUC
rocpred.tree = prediction(pdefault.tree[validsample],truedefault[validsample]) # compute predictions
270:1
(Untitled)
R Script
Console
Jobs
~/Documents/class/data science/examples/lendingclub/
> confmatrix.summary(pdefault.tree[validsample],cdefault.tree[validsample],truedefault[validsample]) # summary for validation sample
$confmatrix
         trueclass
predclass   0   1
        0 2489 382
        1 751 259
$accuracy
[1] 0.7080649
$truepos
[1] 0.4040562
$precision
[1] 0.2564356
$trueneg
[1] 0.7682099
$lift
[1] 1.715072
Environment
History
Connections
Import Dataset
List
Global Environment
cdefault.lr   Large numeric (9578 elements, 673.7 Kb)
cdefault.tree   Large numeric (9578 elements, 673.7 Kb)
evarlist   NULL (empty)
meandata   Named num [1:16] 1 0.122 1.563 0.135 0.413 ...
mvarlist   chr [1:15] "int.rate" "inq.last.6mths" "purposecredit_car…
pdefault.lr   Large numeric (9578 elements, 673.7 Kb)
pdefault.tree   Large numeric (9578 elements, 673.7 Kb)
Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
LogRegr
Tree

Conflicting Instructions from Management

CEO	▪ "Our customers care about one thing – return on their investments. We should make sure that we build a strategy that takes this into account." ▪ "Cost associated with bad loans is killing us (…) the market is penalizing us because they believe that things will only get worse (…) and I also think that they will" ▪ "I think that we made some mistakes during the boom years of the economy (…) we were caught in the origination spiral and gave credits to client that were not credit worthy"
CFO	▪ "Our investors wants us to be profitable, they are not concerned about growing now (….) although if we can provide the required level of return I am sure they will provide additional capital" ▪ "I am sure that we could have increased interest rates but the problem was that everyone was focused on growth (…) I am not even sure if we had the capacity to price correctly our clients" ▪ "With Big Data I can know much better my potential new customers (…) we can price risk much better and generate profitable growth"
CMO	▪ "I think that there is a big opportunity in offering a loans to new clients so we can increase market share (…) we just need to be fast so that we beat Prosper" ▪ "I think there is a lot of potential in creating adaptive interest rates to our existing clients (…) with Internet of Things I can know even how they drive" ▪ "Our existing portfolio is what it is (…) we should be focused on growth the capabilities to not make the same mistakes again"

Examine Prediction Accuracy (by varying cutoff threshold for prediction)

Cutoff = Lower = 0.10

Actual \ Prediction	Default	Don't default
Default	550	91
Don't default	2192	1048

Cutoff = 0.16

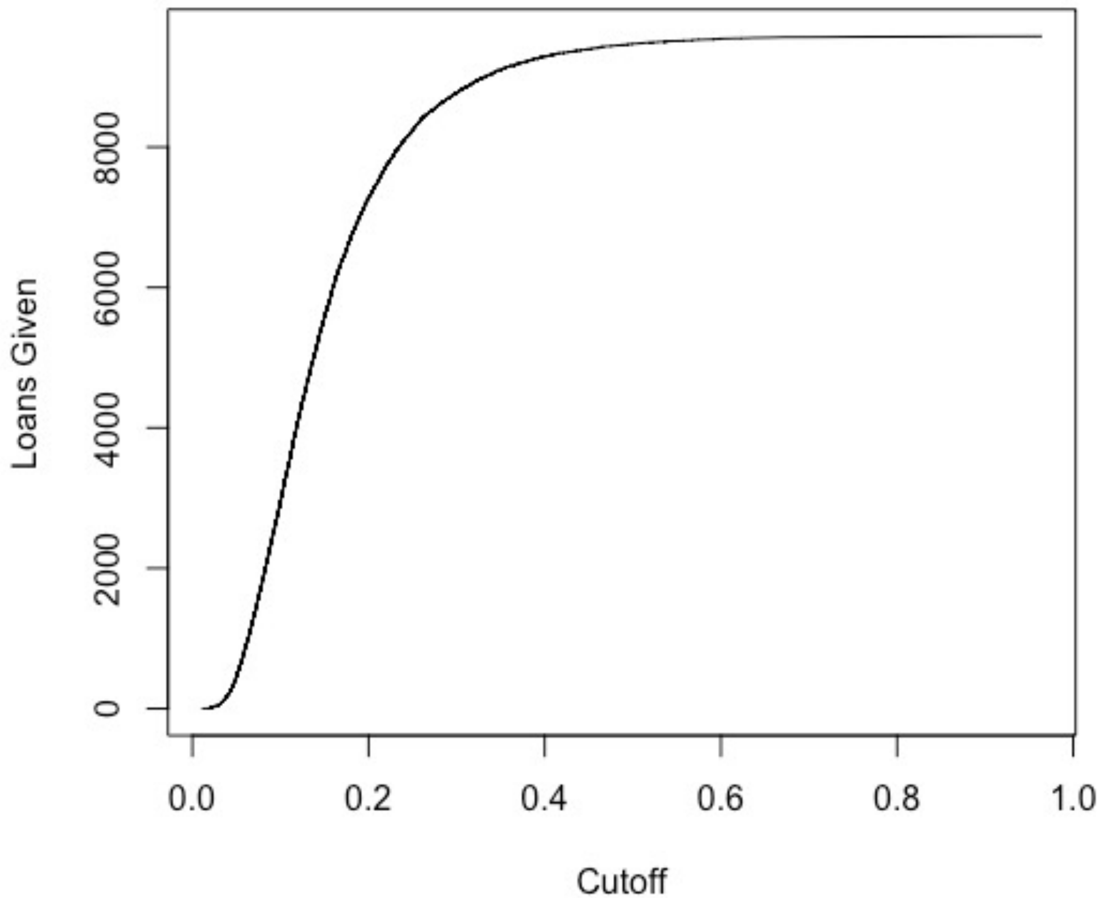
Actual \ Prediction	Default	Don't default
Default	362	279
Don't default	1108	2132

Cutoff = Higher = 0.27

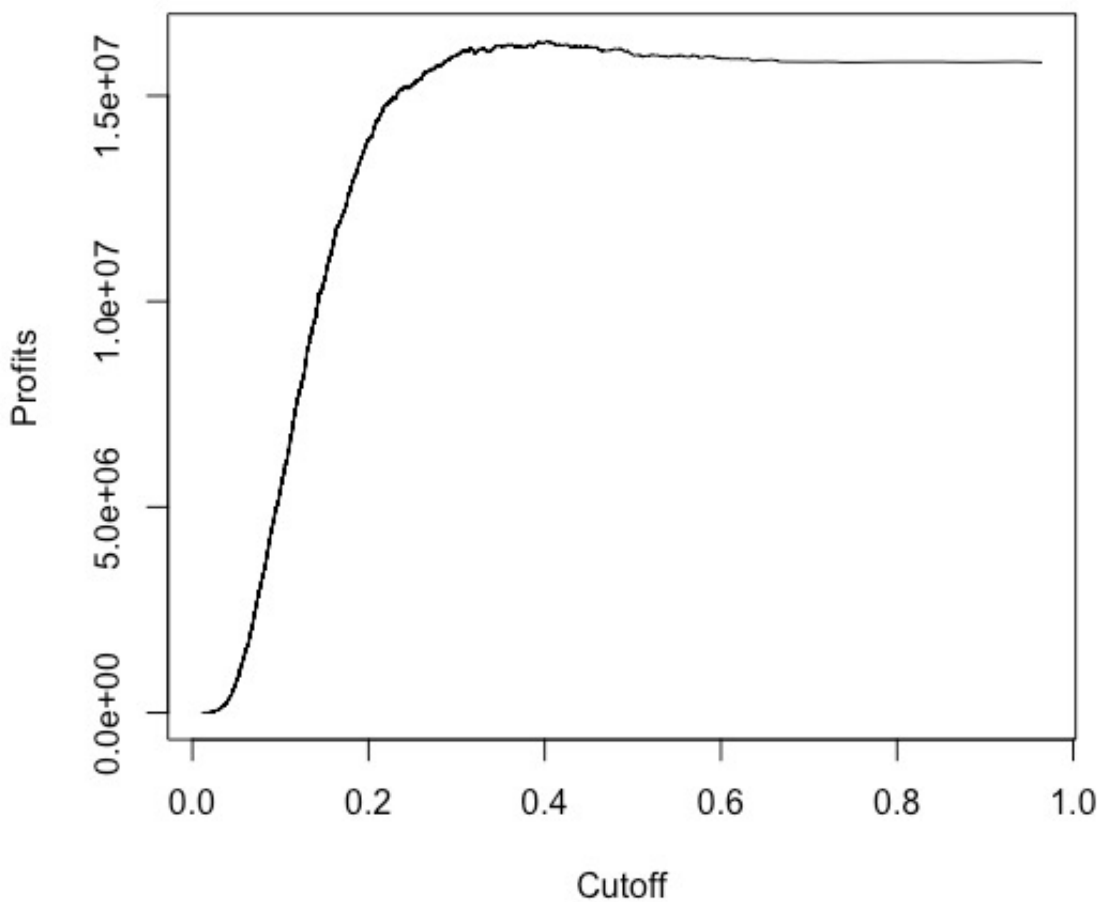
Actual \ Prediction	Default	Don't default
Default	137	504
Don't default	290	2950

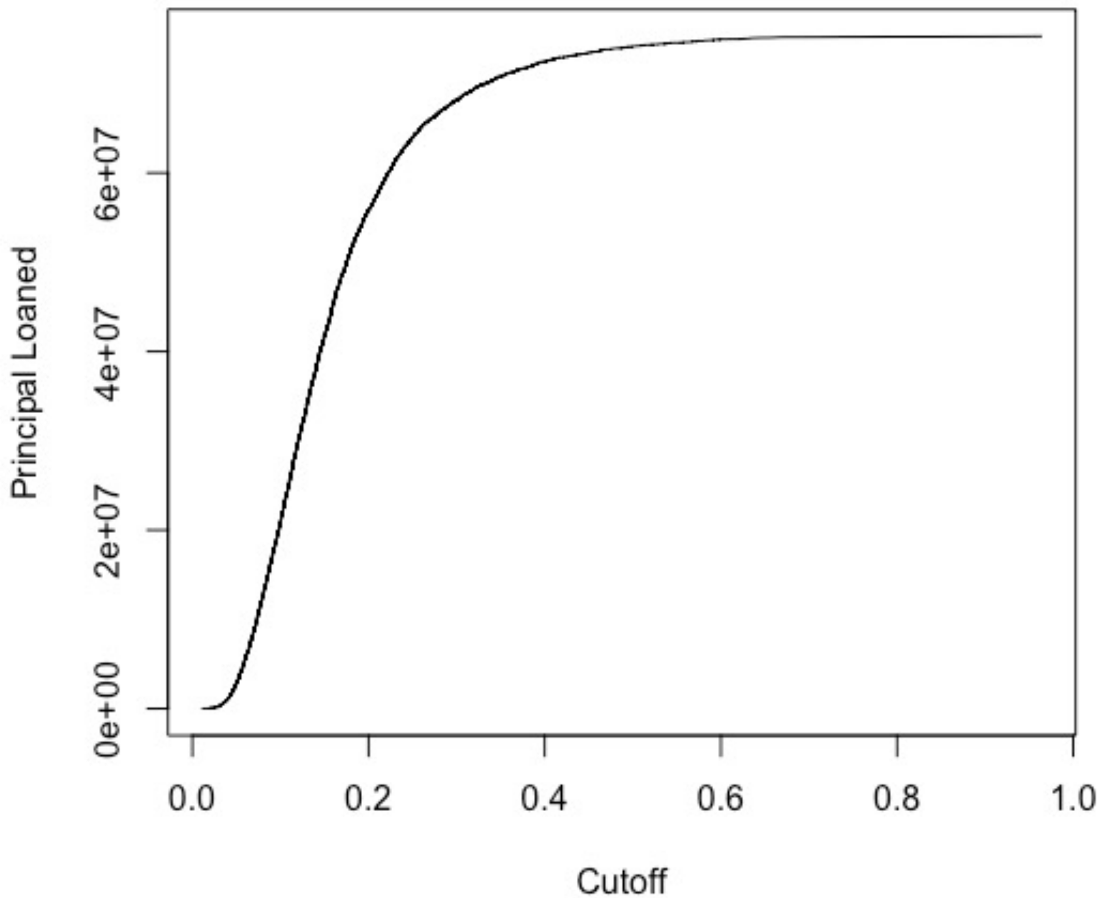
	Cutoff = 0.10	Cutoff = 0.16	Cutoff = 0.27
Loans funded	2944	6014	8506
ROIC	8.7%	8.4%	7.9%
Profit	$5,424,082	$11,380,178	$15,617,018
Loan value	$20,893,580	$45,345,145	$65,988,220

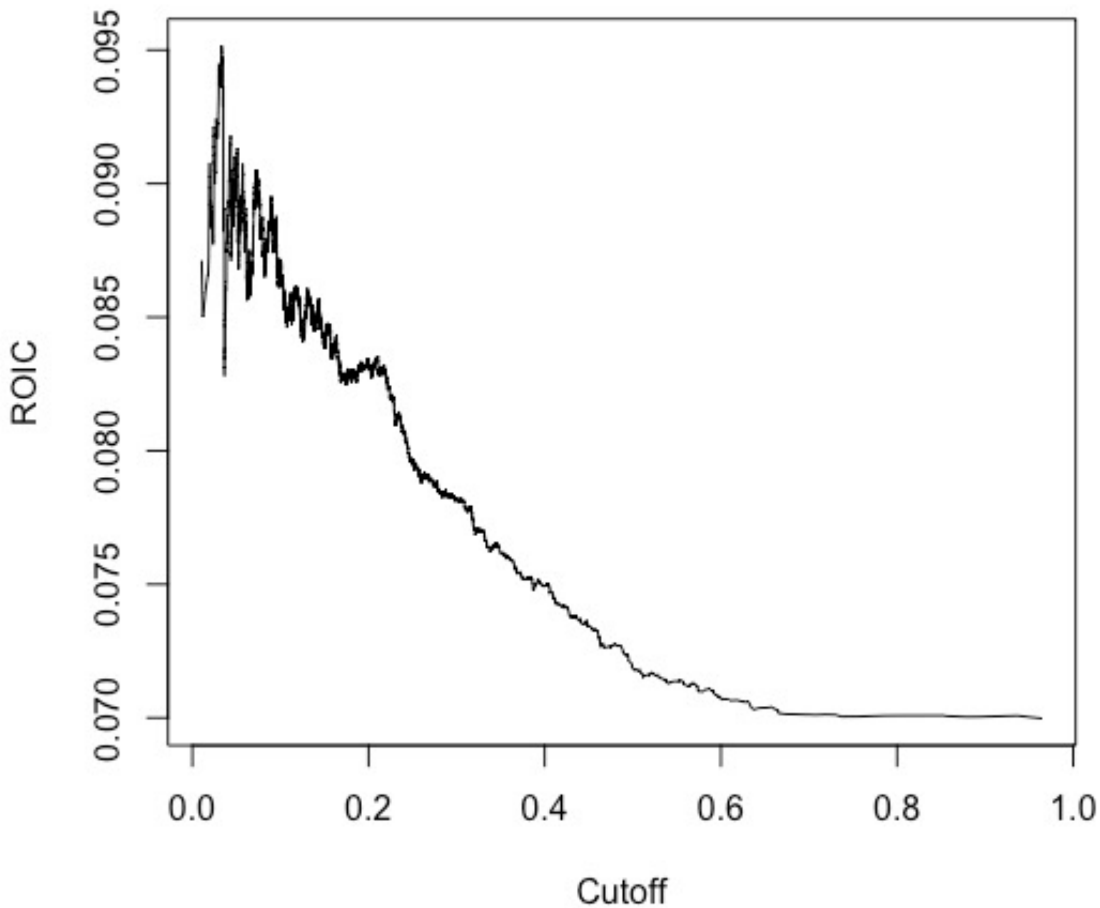
Model choice:
Threshold:



Loans Given
0
2000
4000
6000
8000
0.0
0.2
0.4
0.6
0.8
1.0
Cutoff


Profits
0.0e+00
5.0e+06
1.0e+07
1.5e+07
0.0
0.2
0.4
0.6
0.8
1.0
Cutoff


Principal Loaned
0e+00
2e+07
4e+07
6e+07
0.0
0.2
0.4
0.6
0.8
1.0
Cutoff


ROIC
0.070
0.075
0.080
0.085
0.090
0.095
0.0
0.2
0.4
0.6
0.8
1.0
Cutoff

Summary

Confusion Matrix encapsulates the probability of correct (profitable) and incorrect (unprofitable) decisions

The decision of threshold determines whether we want to be aggressive in making loans (positive classification) or not making loans (negative classification)

Findings

Predictive models are not just about making predictions, but about understanding relationships

You can use predictive models iteratively, to better understand the data, and then (perhaps collect better data and) build better models

Models can be judged on many metrics, but the most important one for a business context is how will it help you improve your decision (and increase profits)