Data Science for Business
Lecture #5
Logistic Regression Example for Lending Club

Prof. Alan L. Montgomery

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

email: alanmontgomery@cmu.edu

LendingClub

Sign in Help

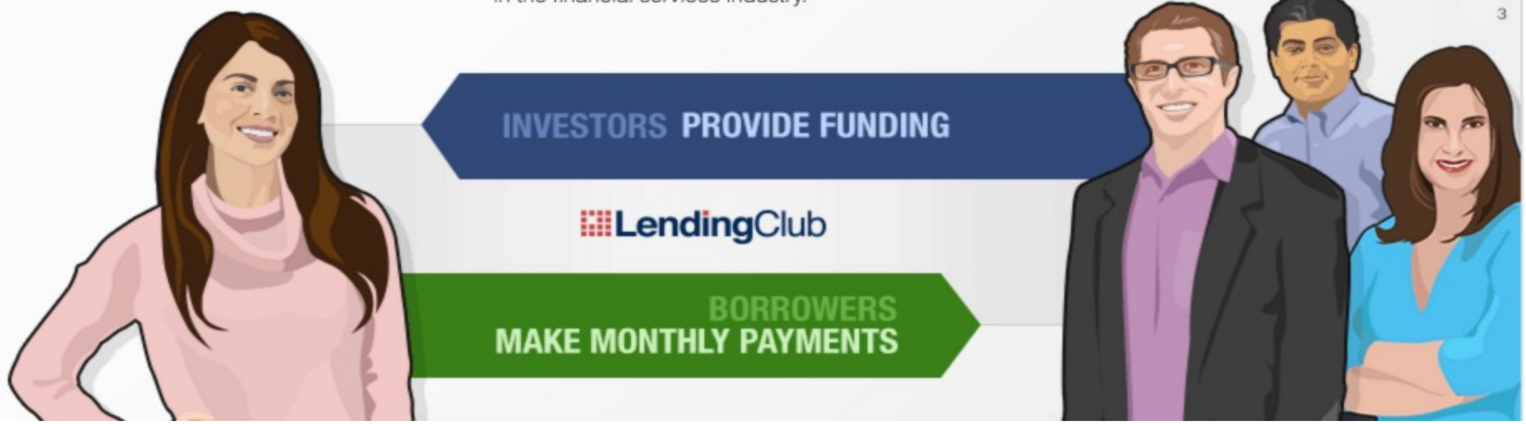
Personal Loans | Auto Refinancing | Business Loans | Patient Solutions | Investing | How It Works | About Us

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]

3



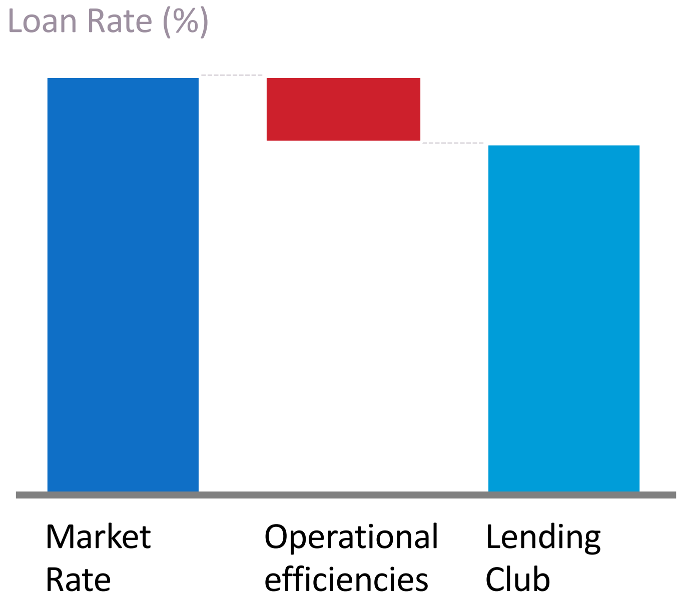
Introduction to Lending Club

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



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).

In-Class Exercise: Part 1
Understand the Data

Review the list of variables and identify what potential relationships you expect to find with loan default

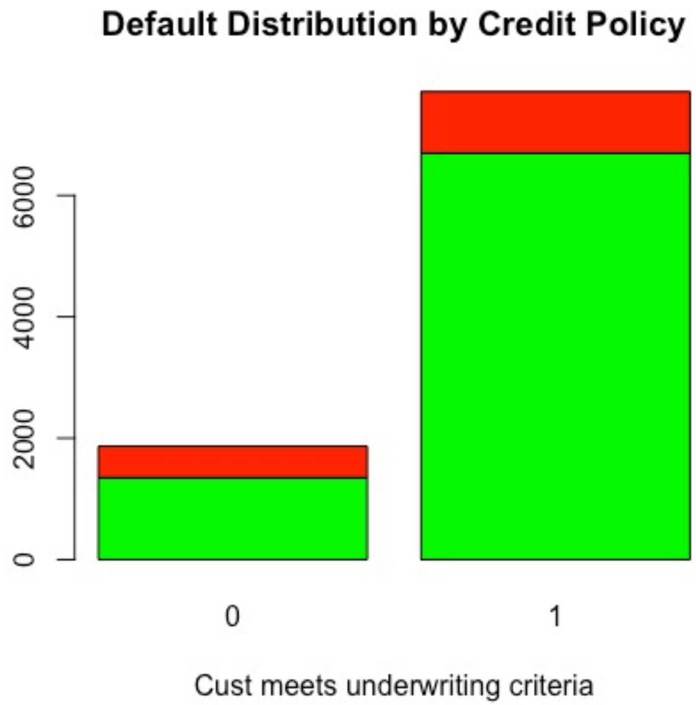
Download the lendingclub_Analysis_logistic.R and loans-default.csv.

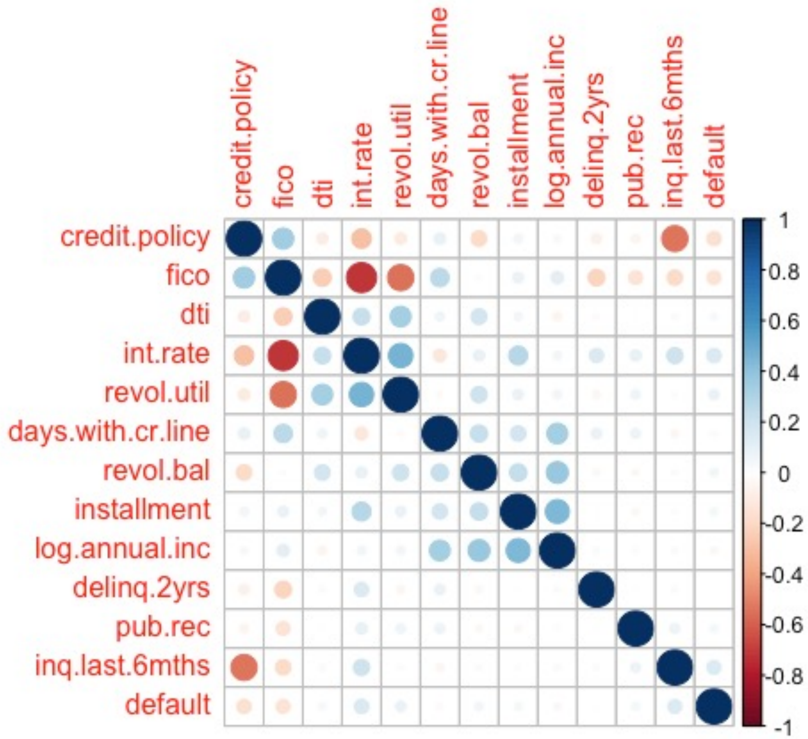
Run the "@setup" and "@input". Step through the "@exploratory analysis" section of the script and carefully consider the output.

What did you learn from descriptive analysis?

What do you expect to learn from a predictive model?

What do we learn from the exploratory analysis?





In-Class Exercise: Part 2
Simple Logistic Regression

Run the “@simple” logistic regression model to estimate a model that predicts *default* using *fico* and *installment*

Simple Logistic Regression

```
> # specify a simple logistic regression model
> lrmdl=glm(default~fico+installment,data=loans[trainsample,],family='binomial')
> # give a summary of the model's trained parameters
> summary(lrmdl)

Call:
glm(formula = default ~ fico + installment, family = "binomial",
    data = loans[trainsample, ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9288  -0.6382  -0.5382  -0.4093   2.5731

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.3691316  0.7488615   8.505  < 2e-16 ***
fico        -0.0118126  0.0010748 -10.991  < 2e-16 ***
installment  0.0008678  0.0001725   5.030  4.9e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4944.4  on 5696  degrees of freedom
Residual deviance: 4797.6  on 5694  degrees of freedom
AIC: 4803.6

Number of Fisher Scoring iterations: 4
```

What is the meaning of this model?

Every 10 point increase in FICO changes the log of the odds ratio of default by (-0.118=10 x -0.0118), and hence the odds ratio by exp(-0.118) = 0.89 (reduces the odds ratio of default by 11%)

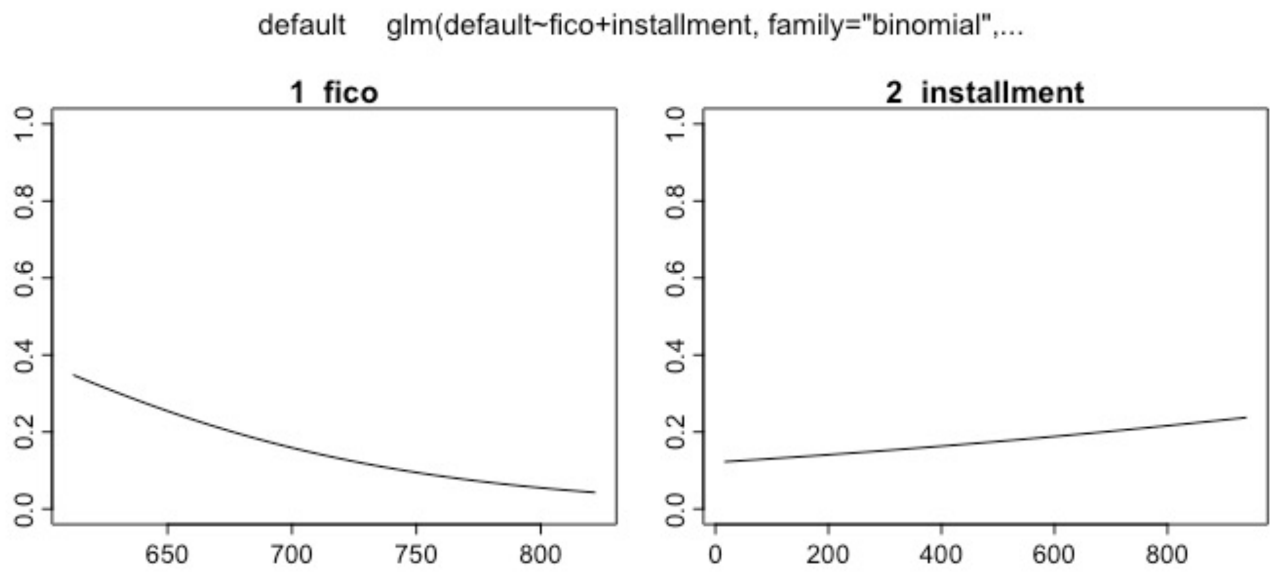
Every $100 increase in installment changes the log-odds ratio of default by (0.087), and hence the odds ratio by exp(0.087) = 1.09 (increases the odds ratio of default by 9%)

Notice that the coefficients depend upon the scale of the variables.

Visualizing Variable impact with plotmo

Plotmo plots the model's response when varying one or two predictors while holding the other predictors constant (e.g., it holds the other values at their median).

It can also generate partial-dependence plots (by specifying pmethod="partdep").



Computing Variable Importance

How much more important is fico than installment?

The answer depends on the coefficient AND the observed variation in the variable

Compute the standard deviation as variation measure

```
> sd(loans$fico)
[1] 37.97054
> sd(loans$installment)
[1] 207.0713
```

To understand the importance compute:

Exp(coef x sd)

In this example:

For FICO: exp(-0.0118 * 37.97) = 0.64

For Installment: exp(0.00087 * 207.07) = 1.20

So which has bigger impact?

In-Class Exercise: Part 3
Logistic Regression

Step through the “@logistic” regression in the script to perform a stepwise logistic regression

What did you learn from the logistic regression about loan default?

Complete the following two slides:

- Explain your model. What variables are important?
- Use your model to construct three different classification matrices

Explaining the Model

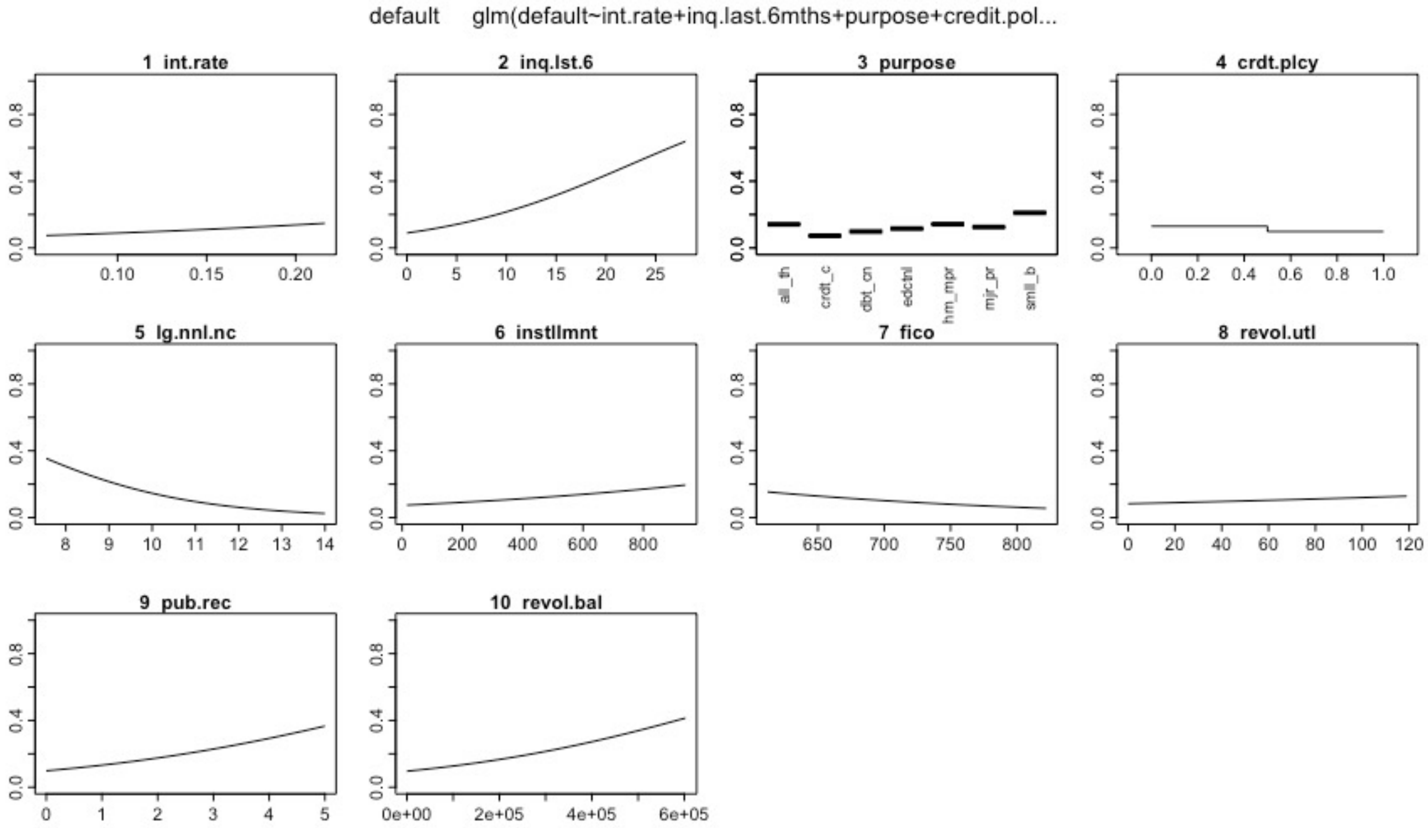
Use this slide to explain your logistic model: do not “copy and paste” estimates – they do not mean anything, instead call out the most important predictors

Explaining the Model

How do we make sense of this model?

```
> summary(lrmdl)

Call:
glm(formula = default ~ int.rate + inq.last.6mths + purpose +
    credit.policy + log.annual.inc + installment + fico + revol.util +
    pub.rec + revol.bal, family = "binomial", data = loans[trainsample,
    ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8004  -0.6054  -0.4805  -0.3523   2.5926

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                6.291e+00  1.603e+00   3.925 8.66e-05 ***
int.rate                   4.907e+00  2.221e+00   2.210  0.02714 *
inq.last.6mths             1.031e-01  1.858e-02   5.549 2.88e-08 ***
purposecredit_card        -7.450e-01  1.476e-01  -5.048 4.47e-07 ***
purposedebt_consolidation -4.150e-01  1.001e-01  -4.148 3.35e-05 ***
purposeeducational        -2.314e-01  2.193e-01  -1.055  0.29126
purposehome_improvement    6.580e-03  1.694e-01   0.039  0.96901
purposemajor_purchase     -1.439e-01  2.001e-01  -0.719  0.47197
purposesmall_business      4.795e-01  1.477e-01   3.246  0.00117 **
credit.policy             -3.170e-01  1.115e-01  -2.842  0.00448 **
log.annual.inc            -4.769e-01  7.534e-02  -6.330 2.45e-10 ***
installment                1.188e-03  2.281e-04   5.210 1.88e-07 ***
fico                      -5.368e-03  1.748e-03  -3.071  0.00214 **
revol.util                 4.083e-03  1.628e-03   2.508  0.01214 *
pub.rec                    3.331e-01  1.197e-01   2.783  0.00539 **
revol.bal                  3.146e-06  1.199e-06   2.625  0.00867 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4944.4  on 5696  degrees of freedom
Residual deviance: 4572.7  on 5681  degrees of freedom
AIC: 4604.7

Number of Fisher Scoring iterations: 5
```

Explaining the Model
Using response plots to understand relationships



Lendingclub_lrmodeldata.csv

	A	B	C	D	E	F	G	H	I	J	K	L
1		rn	Estimate	Std. Error	z value	Pr(>|z|)	meandata	sddata	X1	X2	X3	X4
2	1	(Intercept)	6.2906971	1.60253919	3.92545601	8.66E-05	1	0	1	1	1	1
3	2	int.rate	4.90652818	2.22058796	2.20956264	0.02713553	0.12237816	0.02690847	0.1189	0.1071	0.1357	0.1008
4	3	inq.last.6mth	0.10307346	0.01857638	5.54863041	2.88E-08	1.56310339	2.12725279	0	0	1	1
5	4	purposecredi	-0.7449799	0.14758473	-5.0478113	4.47E-07	0.13533439	0.34211041	NA	NA	NA	NA
6	5	purposedebt_	-0.415028	0.10005148	-4.1481447	3.35E-05	0.41319993	0.49245133	1	0	1	1
7	6	purposeeduc	-0.2314273	0.21928588	-1.0553678	0.29125713	0.03317536	0.17910997	NA	NA	NA	NA
8	7	purposehom	0.00658041	0.16936106	0.03885433	0.96900653	0.06389328	0.24458419	NA	NA	NA	NA
9	8	purposemajc	-0.1439201	0.20009026	-0.719276	0.4719709	0.04598912	0.20947988	NA	NA	NA	NA
10	9	purposesmal	0.4794568	0.14769985	3.24615635	0.00116975	0.06845708	0.2525508	NA	NA	NA	NA
11	10	credit.policy	-0.3169888	0.11153094	-2.8421598	0.0044809	0.80533614	0.39597647	1	1	1	1
12	11	log.annual.in	-0.4768955	0.07533625	-6.3302265	2.45E-10	10.9372354	0.61964747	11.3504065	11.0821426	10.3734912	11.3504065
13	12	installment	0.00118829	0.00022806	5.21038177	1.88E-07	319.784186	207.56305	829.1	228.22	366.86	162.34
14	13	fico	-0.0053679	0.00174805	-3.0707648	0.00213511	711.447077	38.3428417	737	707	682	712
15	14	revol.util	0.00408277	0.00162784	2.50809456	0.01213842	46.3780288	28.8619717	52.1	76.7	25.6	73.2
16	15	pub.rec	0.33314843	0.11971965	2.78273805	0.00539023	0.06371775	0.27214764	0	0	0	0
17	16	revol.bal	3.15E-06	1.20E-06	2.62499277	0.00866508	16879.9665	31533.6528	28854	33623	3511	33667
18	17	pdefault.lr	NA	NA	NA	NA	NA	NA	0.12583553	0.06687281	0.15073092	0.07721207

The lendingclub_Analysis_logistic.R script will output the table of parameter estimates, standard error of the estimates, z- and p-values, as well as the corresponding mean and standard deviation of each of the data variable. The last four columns contain specific users and their predictions.

Compute Variable Importance

	A	B	C	D	E	F	G	H	M	N	O	P	Q
1		Variable	Estimate	Std. Error	z value	Pr(>|z|)	meandata	sddata		Exp(Estimate)	Exp(Est*SD)	1-Exp(Est*SD)	Importance
2	1	(Intercept)	6.29	1.60	3.93	0.00	1.00	0.00		539.53	1.00	0.00	0.00
3	2	int.rate	4.91	2.22	2.21	0.03	0.12	0.03		135.17	1.14	-0.14	0.14
4	3	inq.last.6mths	0.10	0.02	5.55	0.00	1.56	2.13		1.11	1.25	-0.25	0.25
5	4	purposecredit_card	-0.74	0.15	-5.05	0.00	0.14	0.34		0.47	0.78	0.22	0.22
6	5	purposedebt_consolidation	-0.42	0.10	-4.15	0.00	0.41	0.49		0.66	0.82	0.18	0.18
7	6	purposeeducational	-0.23	0.22	-1.06	0.29	0.03	0.18		0.79	0.96	0.04	0.04
8	7	purposehome_improvement	0.01	0.17	0.04	0.97	0.06	0.24		1.01	1.00	0.00	0.00
9	8	purposemajor_purchase	-0.14	0.20	-0.72	0.47	0.05	0.21		0.87	0.97	0.03	0.03
10	9	purposesmall_business	0.48	0.15	3.25	0.00	0.07	0.25		1.62	1.13	-0.13	0.13
11	10	credit.policy	-0.32	0.11	-2.84	0.00	0.81	0.40		0.73	0.88	0.12	0.12
12	11	log.annual.inc	-0.48	0.08	-6.33	0.00	10.94	0.62		0.62	0.74	0.26	0.26
13	12	installment	0.00	0.00	5.21	0.00	319.78	207.56		1.00	1.28	-0.28	0.28
14	13	fico	-0.01	0.00	-3.07	0.00	711.45	38.34		0.99	0.81	0.19	0.19
15	14	revol.util	0.00	0.00	2.51	0.01	46.38	28.86		1.00	1.13	-0.13	0.13
16	15	pub.rec	0.33	0.12	2.78	0.01	0.06	0.27		1.40	1.09	-0.09	0.09
17	16	revol.bal	0.00	0.00	2.62	0.01	16879.97	31533.65		1.00	1.10	-0.10	0.10
18	17	pdefault.lr											

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

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

Cutoff = 0.16

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

```
$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
```

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

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

Cutoff = 0.16

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

Cutoff = Higher = 0.25

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

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

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

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

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)