Introduction to Finance for Data Scientists

Session 6: Scoring

Jean-Edouard Colliard and Johan Hombert

HEC Paris, 2022

Road map

Scoring

- Credit markets

  mortgages, consumer loans, business loans

  $\Rightarrow$ lender must predict default

- Insurance markets

  health, property and casualty

  $\Rightarrow$ insurer must predict losses and damages

- Accurate prediction is key

  - Price too low $\Rightarrow$ lose money
    Price too high $\Rightarrow$ lose market shares
  - This morning: predict default when borrower has a stock price. Not applicable to private companies and people
  - Now, scoring: use data to predict default

Credit scoring using alternative data

- Credit score providers: use data on people and businesses to calculate credit scores sold to banks

"Social medial insight program that extracts data from Yelp, Facebook, Twitter, and Four Square is offered for use by private lenders and traditional banks. Credit scores for over 1 billion people & businesses, including 235 individuals US consumers and over 25 million US businesses."

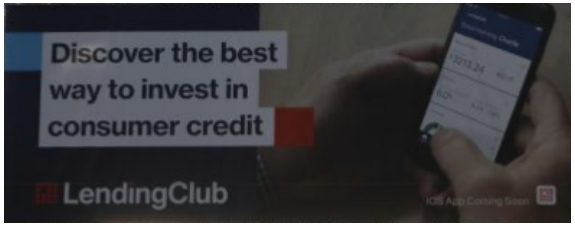
Also done internally by banks

P2P lending (a brief history)

PROSPER From crowdlending to marketplace lender

- Crowdlending: Match borrowers and lenders, terms of financing set by lenders
- Marketplace: Score borrowers, set the interest rate, match borrowers with lenders

LendingClub From marketplace to (shadow) bank



LendingClub Closing Down Their Platform for Retail Investors

Peter Renton · Peer to Peer Lending · Oct. 7, 2020 · 5 min read

- (Shadow) Bank: Score borrowers, set the interest rate, lend using its balance sheet

Big tech in credit markets

- Tech giants are well positioned in credit markets: they have access to consumers + have lots of data on them

THE WALL STREET JOURNAL.

TECH

Goldman Sachs, Apple Team Up on New Credit Card

Card would carry the Apple Pay brand and could launch early next year

Apple buys UK fintech start-up Credit Kudos

Purchase suggests US tech giant will launch greater push into credit services

Tim Bradshaw and **Siddharth Venkataramakrishnan** in London MARCH 23 2022

InsurTech

- Creation of insurance products, scoring with AI

Data machine: the insurers using AI to reshape the industry

Groups are building detailed customer profiles to inform pricing and try to influence behaviour

AI allows insurers such as Ping An to produce highly individualised profiles of customer risk that evolve in real time © FT montage; Alamy, Dreamstime

Ian Smith in London AUGUST 16 2021

50

Case study

- "On the Rise of FinTechs: Credit Scoring Using Digital Footprints," 2019, Berg, Burg, Gombovic and Puri, *Review of Financial Studies* [pdf]

– Scoring with digital footprints at an e-commerce company

  – Goods sent first, paid for later → need to assess buyer's creditworthiness

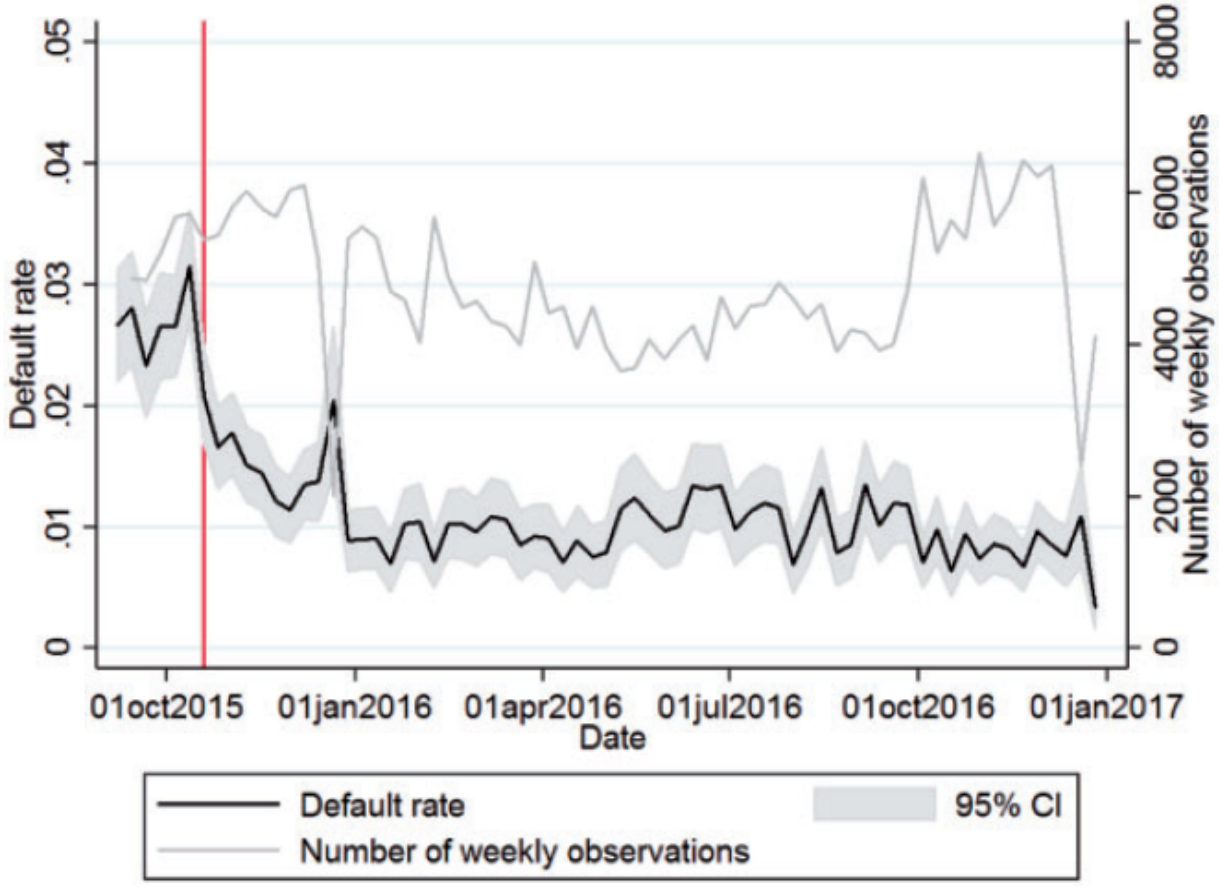
  – Credit score based on credit history, sociodemographics, past transactions, etc.

  – After Oct 2015: also collected digital footprints (OS, email, etc.)

  – Does this improve prediction of default?

Default regressions (scorable customers)

	(1) Credit bureau bureau score		(2) Digital footprint		(3) Credit bureau score & digital footprint	
Variables	Coef.	z-stat	Coef.	z-stat	Coef.	z-stat
Credit bureau score	−0.17***	(−7.89)			−0.15***	(−6.67)
Device type & operating system[a]						
Desktop/Windows			Baseline		Baseline	
Desktop/Macintosh			−0.07	(−0.53)	−0.13	(−1.03)
Tablet/Android			0.29***	(3.19)	0.29***	(3.06)
Tablet/iOS			0.08	(1.05)	0.08	(0.97)
Mobile/Android			1.05***	(17.25)	0.95***	(15.34)
Mobile/iOS			0.72***	(9.07)	0.57***	(6.73)
E-mail Host [a]						
Gmx (partly paid)			Baseline		Baseline	
Web (partly paid)			0.00	(0.00)	−0.02	(−0.22)
T-Online (affluent customers)			−0.40***	(−3.90)	−0.35***	(−3.35)
Gmail (free)			0.34***	(3.81)	0.29***	(3.09)
Yahoo (free, older service)			0.75***	(9.19)	0.72***	(8.98)
Hotmail (free, older service)			0.35***	(3.70)	0.28***	(2.72)
Channel						
Paid			Baseline		Baseline	
Affiliate			−0.49***	(−5.35)	−0.54***	(−5.58)
Direct			−0.27***	(−4.25)	−0.28***	(−4.44)
Organic			−0.15*	(−1.79)	−0.15*	(−1.74)
Other			−0.47***	(−4.50)	−0.48***	(−4.36)
Checkout time						
Evening (6 p.m.-midnight)			Baseline		Baseline	
Morning (6 a.m.-noon)			0.28***	(4.50)	0.28***	(4.60)
Afternoon (noon-6 p.m.)			0.08	(1.42)	0.08	(1.47)
Night (midnight-6 a.m.)			0.79***	(7.73)	0.75***	(7.09)
Do-not-track setting			−0.02	(−0.25)	−0.07	(−0.91)
Name in e-mail			−0.28***	(−5.67)	−0.29***	(−5.70)
Number in e-mail			0.26***	(4.50)	0.23***	(3.91)
Is lowercase			0.76***	(13.10)	0.74***	(13.20)
E-mail error			1.66***	(20.00)	1.67***	(20.36)
Constant	12.42***	(5.76)	−4.92***	(−62.87)	9.97***	(4.48)
Control for *Age*, *Gender*, *Item category*, *Loan amount*, and month and region fixed effects	No		No		No	
Observations	254,819		254,819		254,819	
Pseudo R^2	**.0244**		**.0524**		**.0717**	
AUC	**0.683**		**0.696**		**0.736**	
(SE)	**(0.006)**		**(0.006)**		**(0.005)**	
Difference to AUC=50%	**0.183*** **		**0.196*** **		**0.236*** **	
Difference AUC to (1)			**0.013***		**0.053*** **	


Default rate
0
.01
.02
.03
.04
.05
Number of weekly observations
0
2000
4000
6000
8000
01oct2015
01jan2016
01apr2016
01jul2016
01oct2016
01jan2017
Date
Default rate
95% CI
Number of weekly observations

Issues for data scientists

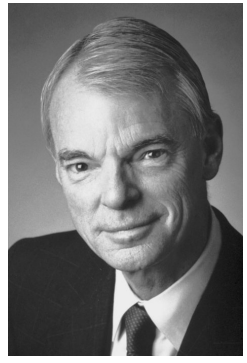
- Scoring with AI can be very powerful
- But also important pitfalls to avoid → today's lecture

Road map

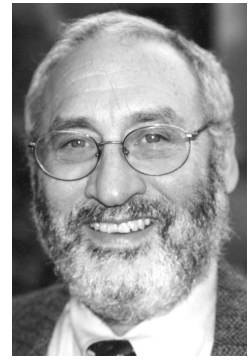
Adverse selection

which took a few Nobel prizes to figure it out...

George A. Akerlof (2001)

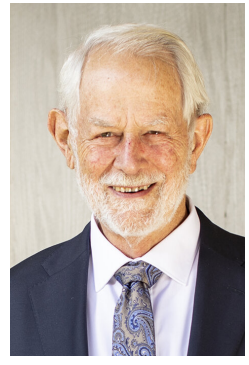


A. Michael Spence (2001)



Joseph E. Stiglitz (2001)

Paul R. Milgrom (2020)



Robert B. Wilson (2020)

Fintech lender

- Example: Fintech lender making loans to businesses
  - Receive loan applications
  - Info on loan applicants: vector X (financial info, online reviews, ...)
  - Offer interest rate R. Loan applicant may take the loan or not
  - If loan is taken, then cash flow is

    today: $-1 + R$

    at maturity: $\begin{cases} 1 \text{ if no default} \\ 0 \text{ if default} \end{cases} = 1 - D$ where $D \in \{0, 1\}$ is default indicator

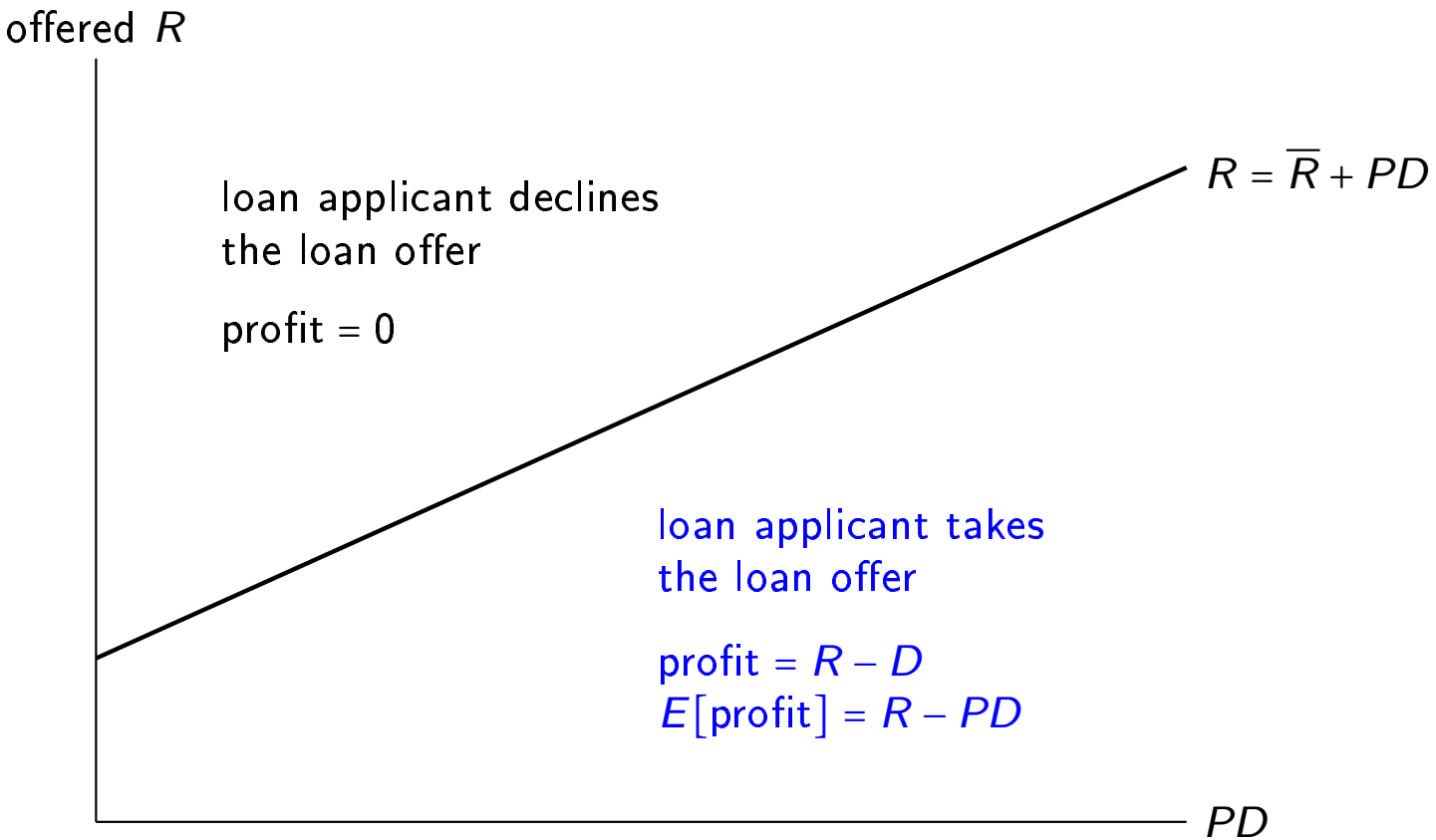
    (Unimportant assumptions to simplify formulas: interest paid at issuance; loan size normalized to one; zero recovery rate; zero discount rate)

  $\Rightarrow$ Profit $= 1\{\text{loan taken}\} \times (R - D)$

Adverse selection

- Probability of default "PD" = $f(X) + U$
  - X: observed characteristics
  - U: unobserved determinants of default, uncorrelated with X, $E[U] = 0$
- Borrower offered an interest rate R takes the loan iff $R < \overline{R} + PD$
  - Borrowers take the loan if the rate is not too high
  - $\overline{R}$: maximum rate a risk-free borrower would accept
  - Riskier borrowers are more likely to take the loan (because their chance to get a loan from another bank are low)

Adverse selection



Adverse selection

- Predicting default

Step 1: Construct scoring model. You have data on past loans with info (X, D). You recover $f(.)$ using ML

Step 2: Score new applications. You have info X on new applications. Your best estimate of PD is $P[D|X] = f(X)$

- How to set the interest rate?

- Simple idea: take a margin over estimated PD: $R = f(X) + M$ with $M > 0$

Q. What is your expected profit per loan granted?

a. M b. more than M c. less than M d. it depends

Adverse selection

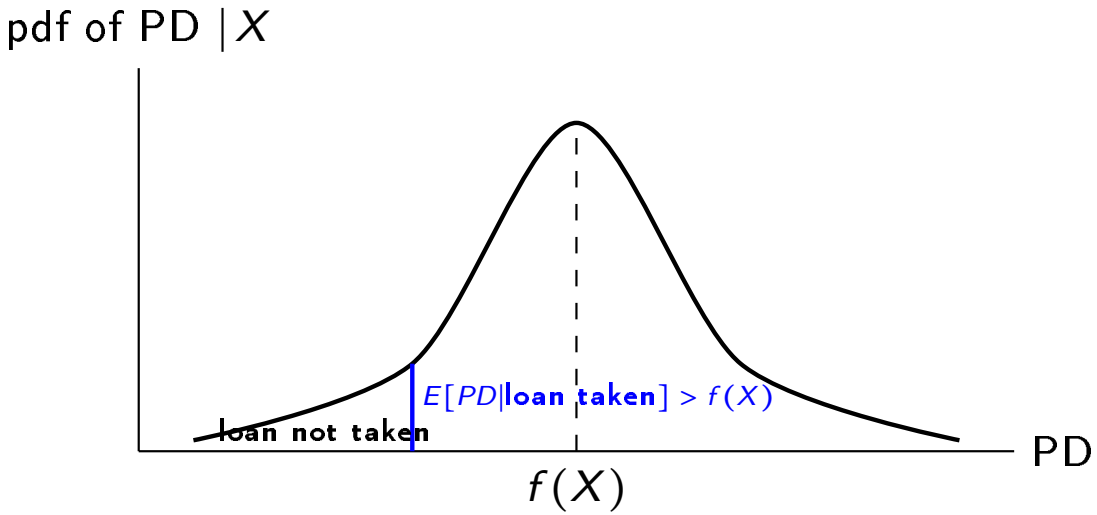
- Expected profit per loan granted (conditional on X)

$$
\begin{aligned}
&= E[R - D \,|\, \text{loan is taken}] \\
&= R - E[D \,|\, R < \overline{R} + PD] \\
&= f(X) + M - E[f(X) + U \,|\, f(X) + M < \overline{R} + f(X) + U] \\
&= M - E[U \,|\, U > M - \overline{R}] \quad < \quad M \quad !!!
\end{aligned}
$$

$\Rightarrow$ You earn less than M per loan. What happened?

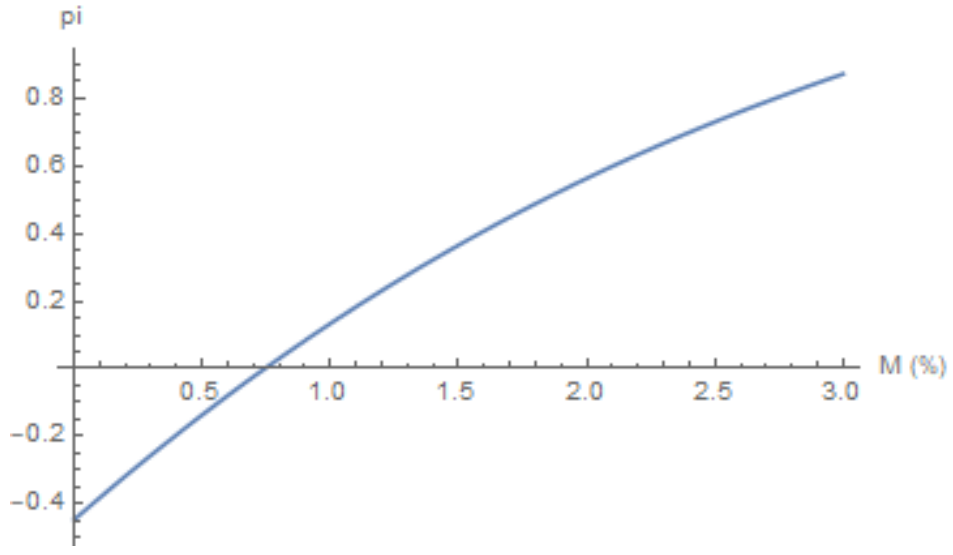
Adverse selection

- **Adverse selection:** The pool of applicants who accept the loan offer are more risky than the overall pool of applicants (i.e., the pool of accepted offers is *adversely selected*)



Adverse selection

- Expected profit per loan granted as function of M

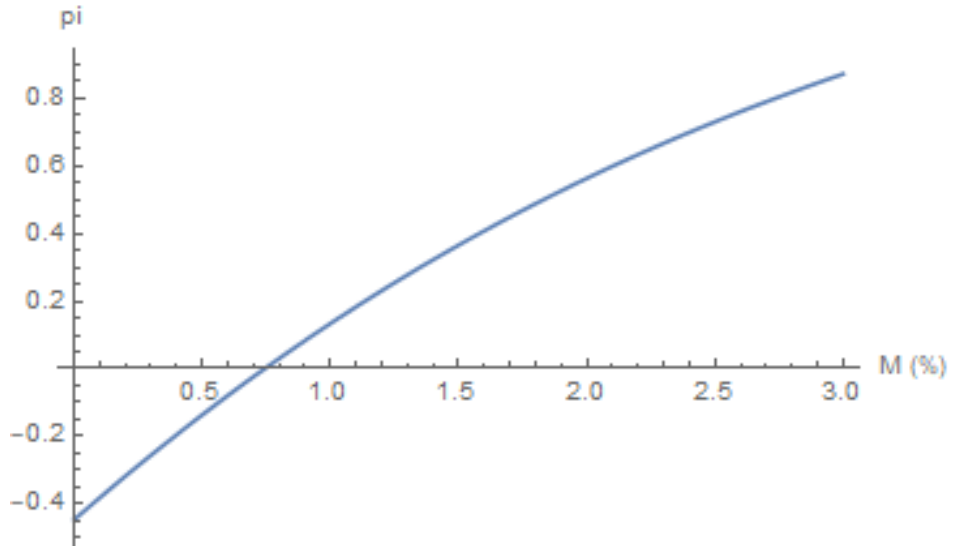


$\overline{R} = 2\%$; $U \sim \mathcal{N}(0, 1.8\%)$

- Adverse selection $\Rightarrow$ profit per loan less than M; can be negative even if $M > 0$

Adverse selection

- Expected profit per loan granted as function of M



$\overline{R} = 2\%$; $U \sim \mathcal{N}(0, 1.8\%)$

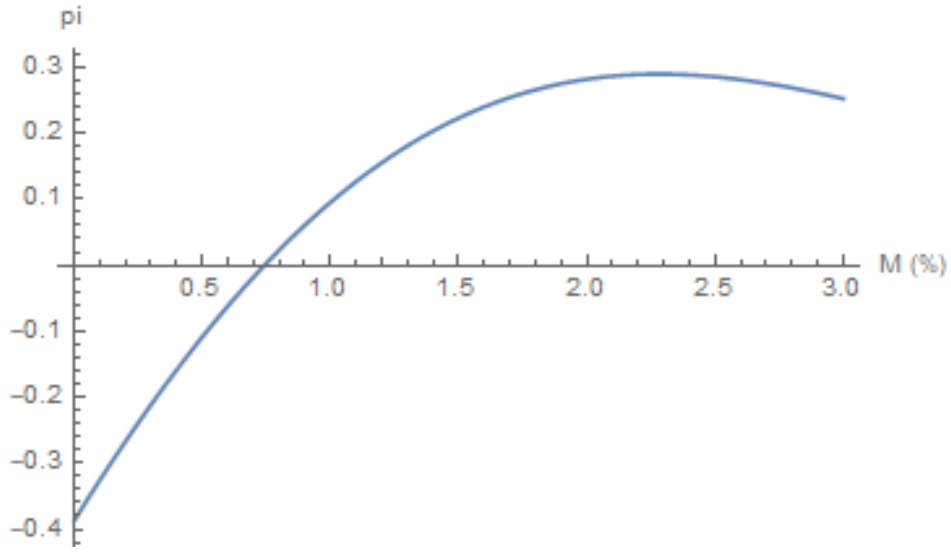
- Adverse selection $\Rightarrow$ profit per loan less than M; can be negative even if $M > 0$

- Next question: Which M maximizes expected profit?

Adverse selection

- Expected profit

$$= E[(R-D)\times 1\{\text{loan taken}\}] = \underbrace{E[R-D\,|\,\text{loan taken}]}_{\uparrow M \text{ (previous graph)}} \times \underbrace{P[\text{loan taken}]}_{\downarrow M}$$



- Why humped-shaped? Higher $M \Rightarrow$ higher profit per loan granted but fewer loan offers are accepted

- Profit is maximized for $M \approx 2.2\%$

Adverse selection — Summary

- Adverse selection is similar to the "winner's curse" (as in the auction you played with Jean-Edouard)

- If (and only if) competitors or customers have info about default/losses that you don't have, customers that take your offer are worse than average

Adverse selection — What should I do?

- Ask yourself which information YOU DON'T HAVE and others have

  If that information is correlated with the variable to predict, you face an adverse selection problem

- Practical solution: adjust your forecast (be more conservative)

- Even better: A/B testing on the interest rate

Road map

Selection bias

- Previous example: data on past loans are unbiased → you can recover $f(.)$ without bias
- Selection can also bias past data → problematic to recover $f(.)$
- Example
  - OldStyleBank is a traditional bank that makes loans to small businesses. It employs loan officers who meet with entrepreneurs before deciding to make loans. It asks you to develop a credit scoring model to replace loan officers
  - You have data on past loans: borrower characteristics X, default indicator D
  - Goal: construct a scoring model which estimates the probability of default $P[D|X]$

Selection bias

- True probability of default is $f(X) + U$

  X: observed characteristics (financial statements, etc.)

  U: characteristics you don't observe but that the bank loan officer observed (e.g., entrepreneur is hard-working and trustworthy), uncorrelated with X, $E(U) = 0$

  Suppose loan officers granted the loan iff $U < \overline{U}$

- Simple idea: Estimate $P[D \mid X] = f(X)$ as the default rate conditional on X in past loans data

Q. Is it an unbiased estimate of the default probability for future loan applications? (assuming future loan applications are drawn from the same population as past ones and that there is no adverse selection)

a.yes b. it under-estimates true default proba c. it over-estimate true default proba

Selection bias

- Default rate conditional on X in past loans data

$$P[D \mid X, \text{in past data}] \;=\; f(X) + E[U \mid U < \overline{U}] \;<\; f(X) \text{ !!!}$$

$\Rightarrow$ Default rate of loans granted $<$ default probability of loan applicants

Why?

- **Selection bias:** Loans appear in the data only if the loan officer had positive info on the borrower that you don't have

- If you don't take this selection bias into account, you under-estimate the true default probability. If you set the interest rate based on this estimate, you lose money

Selection bias — What should I do?

- Very important to know HOW DATA WERE SELECTED. Most of the time not random

- Ask yourself whether data selection is based on information 1) you don't have and 2) is correlated with the variable to predict

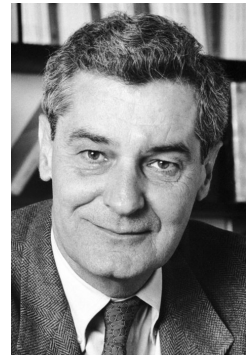
  If yes, there is a selection bias problem

- Solution: ask yourself in which direction selection tilts the distribution of the variable to predict in the data. Adjust your score in the opposite direction

Road map

Lucas critique

which took another Nobel prize

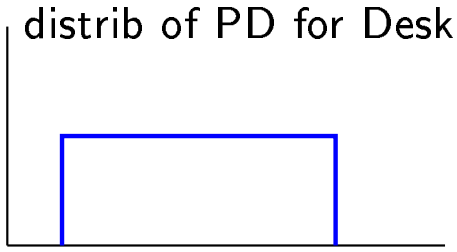


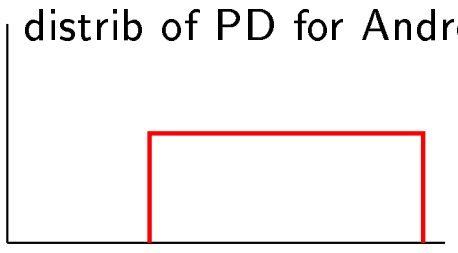
Robert E. Lucas Jr. (1995)

- Basic idea: People adapt their behavior to companies/governments' policies

Lucas critique

- Suppose lender observes whether borrowers connect from Android or Desktop and did not use this info in the past





- Lender starts setting lower rate for Desktop users and higher rates for Android users. Some Android users find out and switch to Desktop

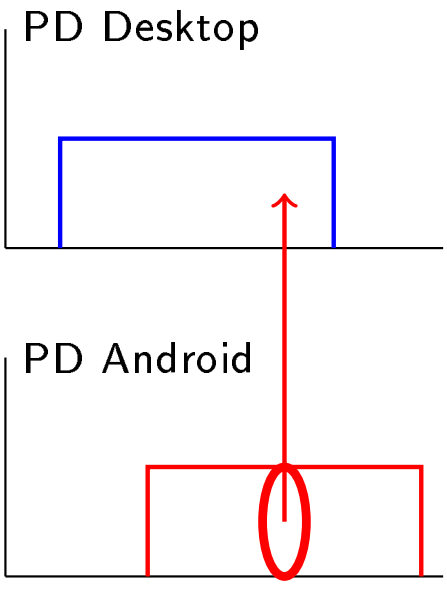
Q. How does this affect the average PD of Desktop users?

a. increases b. decreases c. may increase or decrease

Lucas critique

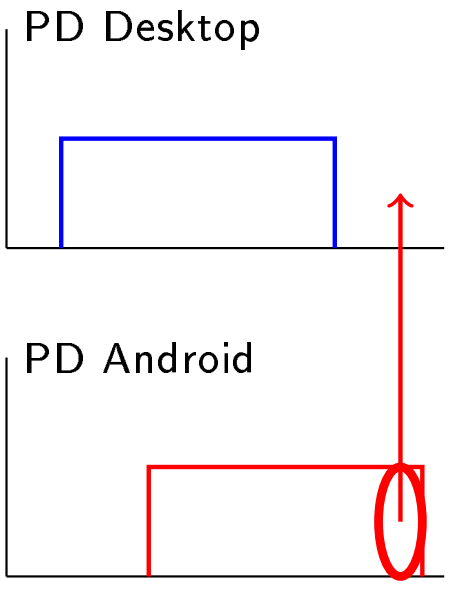
- It depends on who are the switchers!

If intermediate Android users switch



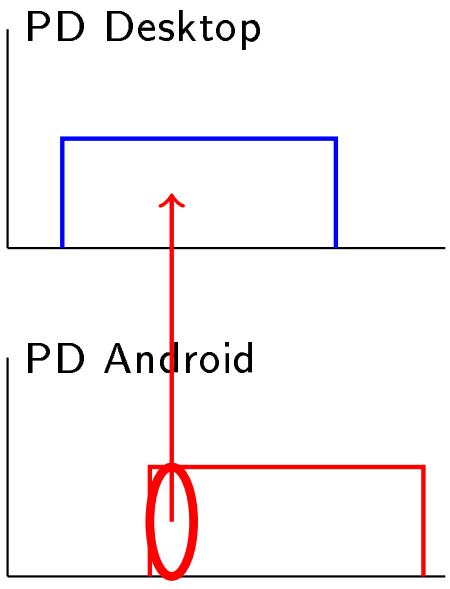
PD of Desktop users ↑

If risky Android users switch



PD of Desktop users ↑

If safe Android users switch

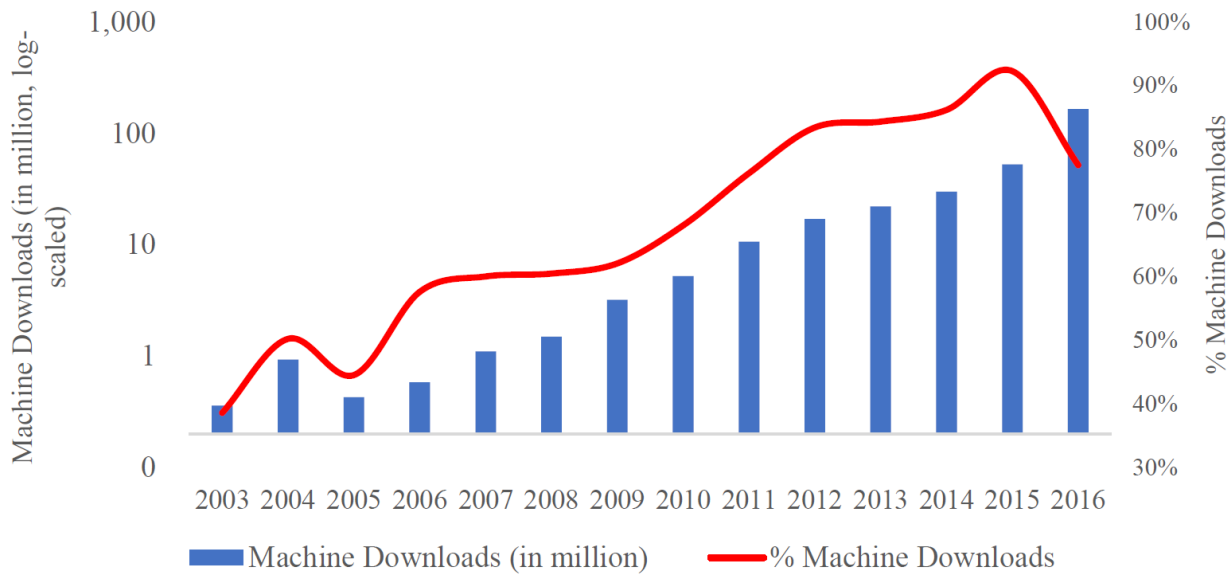


PD of Desktop users ↓

Case study: Corporate reporting

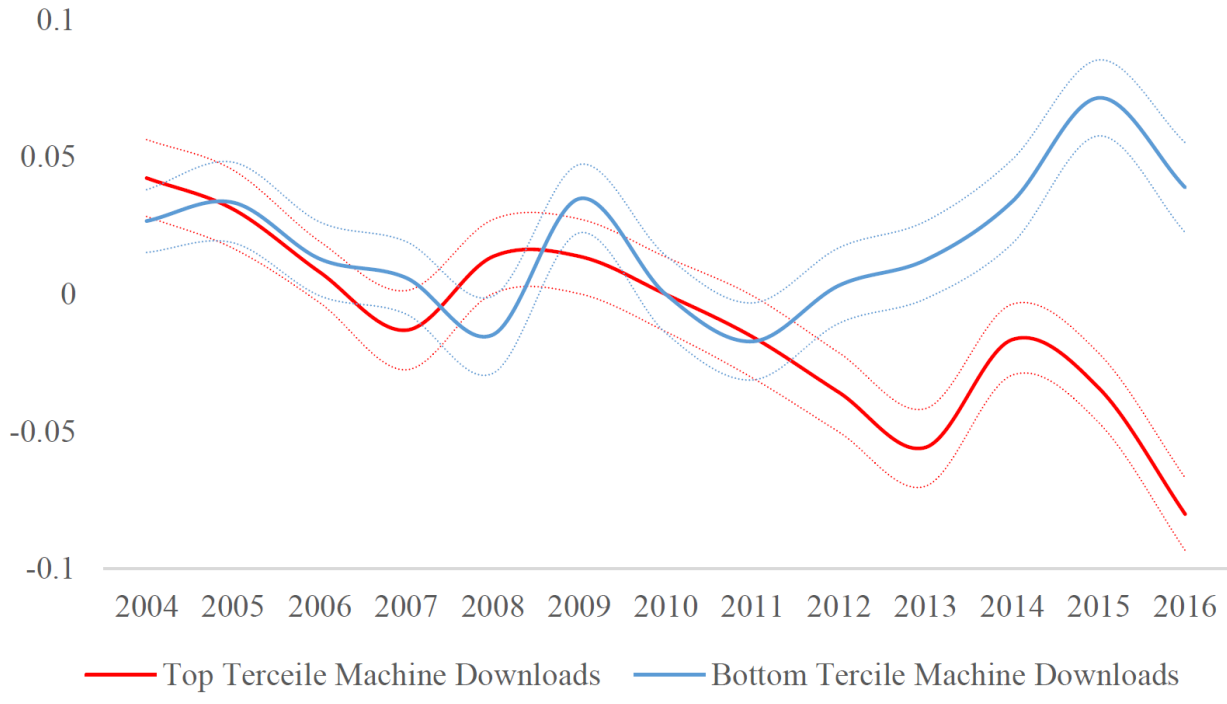
- "How to Talk When a Machine is Listening: Corporate Disclosure in the Age of AI," Cao, Jiang, Yang and Zhang, 2020 [pdf]

Machine downloads of companies' filings with the SEC



Case study: Corporate reporting

- Use of negative words as measured by Loughran-McDonald dictionary widely used by quant funds → decreases after post-2010 rise of algos for companies scrutinized by algos (in red)



Lucas critique: Implications

1. Beware scoring on behavior that can strategically be modified

   $\Rightarrow$ Ask yourself if data is exogenous or the outcome of a strategic choice

2. A good predictor when not used, can become a poor predictor once used (Goodhart's law)

   $\Rightarrow$ Check if the predictive power changes after data is used

Lucas critique: Implications

3. More data can sometimes make everyone (companies AND consumers) worse off

   – Android mobile owners must connect from a desktop or buy a new mobile → they're worse off

   AND the predictive power of Android has disappeared → fintech is worse off

   – Companies avoid certain words in their filings, creating unnecessary complexity for everyone

Lucas critique: Implications

4. Commitment NOT TO use data can sometimes create value

   - Privacy can be a source of economic value

   - Customers' valuation for privacy can be estimated using A/B testing

"The Value of Privacy: Evidence from Online Borrowers"
Huan Tang (HEC PhD 2020) [pdf]

   - Credibility of commitment is key: once personal data exist, it is tempting for companies to use them

Road map

Information in insurance

- Suppose we discover how to predict perfectly who will get sick (but this foreknowledge does not help to prevent or cure diseases)

  Insurers use this information to price health insurance

Q1. Will this make people better or less-well insured?

a. better insured b. less-well insured

Q2. Will this make insurers more or less profitable?

a. more profitable b. less profitable

Information in insurance

- What will happen?
  - No insurers accept to insure people predicted to be sick
  - People predicted to be healthy don't need insurance

$\Rightarrow$ The health insurance market breaks down

  - People are worse off: they can't get insurance
  - Insurers are worse off: they can't sell insurance

- This is the **Hirshleifer effect**: Information can destroy insurance

How to overcome the Hirshleifer effect?

- Suppose an insurer announces it will not use the information (and suppose it is credible)

Q3. Does it overcome the Hirshleifer effect and allow the insurer to sell insurance?

a. yes b. no

How to overcome the Hirshleifer effect?

- Suppose an insurer announces it will not use the information (and suppose it is credible)

Q3. Does it overcome the Hirshleifer effect and allow the insurer to sell insurance?

a. yes b. no

- No, because of adverse selection
  - People predicted to be healthy are offered cheap insurance from other insurers (or they don't even buy insurance)
  - The insurer only gets people who will be sick, so it cannot insure them

How to overcome the Hirshleifer effect?

- Solution 1: Ensure no insurer uses the information
  - Industry ethical standard?
  - Regulation (e.g., insurers are prohibited from using genetic information)

How to overcome the Hirshleifer effect?

- Solution 2: Insure before information is revealed
  - Long-term insurance / Premium guaranteed over long period

Road map

Direct discrimination

- Direct discrimination: Decision based on a "protected characteristic" such as race, sex, ethnic origin, social origin, religion
  - Called "disparate treatment" in US law
- May happen because of prejudice
  - Ex.: Job opening for white men only
- . . . or because the protected characteristic is a predictor of risk (a.k.a. statistical discrimination)
  - Ex.: Cheaper car insurance for women because they have fewer accidents
- Illegal in both cases in EU and US

Indirect discrimination

- Indirect discrimination: Decision is not based on protected characteristics but ends up being different for people with a protected characteristic
  - Called "disparate impact" in US law
- Happens when decision is based on variables correlated with protected characteristics
  - Ex.: Interest rate based on borrower's job occupation may end up being different across people with different ethnic origins
- May be legal or illegal (legal for "business necessity" in US law)

Data and discrimination

The New York Times

Apple Card Investigated After Gender Discrimination Complaints

A prominent software developer said on Twitter that the credit card was “sexist” against women applying for credit.

Data and discrimination

Wells Fargo, Upstart criticized after study finds loan disparities

Published Feb. 6, 2020 • Updated Feb. 14, 2020

The request comes a week after the nonprofit Student Borrower Protection Center found that an Upstart borrower who attended historically black Howard University would pay thousands of dollars more on average for a five-year loan than a borrower with an identical credit profile who studied at New York University.

Discriminatory algorithms?

- Algos are (a priori) not subject to human prejudices but...

1. Biased data: algos are fed with human-world data, which may be contaminated by discrimination

2. Triangulation: algos may "triangulate" protected characteristics from other data (without intent to do so) and use it

- A solution: algorithm interpretability, understand how algos make decisions

Case study: Fintech lenders in the US mortgage market

- "Consumer Lending Discrimination in the FinTech Era," Bartlett, Morse, Stanton and Wallace, *Journal of Financial Economics*, 2021 [pdf]

- For given borrower characteristics, Latin and African-American mortgage borrowers pay higher interest rates

  +8 basis points per year at traditional lender

  +5 basis points per year at fintech lender

Case study: Fintech lenders in the US mortgage market

- Half full glass
  - Less discrimination by fintech
  - Discrimination by traditional lender has decreased over time, perhaps as a result of competition from fintech
- Half empty glass
  - Algorithms still discriminate (although less so than humans)
  - Algorithms "learn" that Latin/African-American borrowers are less likely to get a good rate from a traditional lender, so they can be charged higher rates

Group work — Lending game

- 3 fintech lenders (=3 teams of students) compete to make loans
- Data on past loans: predictors + default
- New loan applications
- All teams have data on the same past loans and receive the same new loan applications, but each team has different predictors
- The game (detailed guidelines on Slack)

Stage 1: Each team makes an offer to every loan applicant → offers due on October 8

Each loan applicant chooses which team's loan offer to take.
Applicants repay or default → profits and losses for each team

Stage 2: Each team improves its strategy based on the experience of stage 1 and play the same game again → offers + report due on October 14

- If not already done: send me your team's composition today

Conclusion

- I hope we conveyed that finance is fun and full of useful ideas ☺

- More advanced finance courses this year
  - Data Analysis in Finance
  - Economic Value of Data
  - Cryptocurrencies and Blockchains

Thank You!