Hi! PARIS Summer School 2022

Tutorial 1A

Data in Finance: FinTech Lending

Johan Hombert (HEC Paris)

July 4, 2022, 11:00-15:00

Road map

What is finance?

Lending game

[optional] AI and discrimination

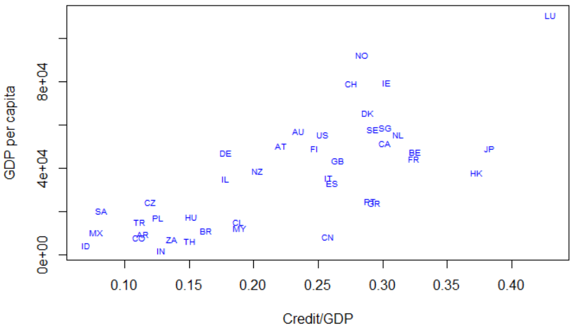
Conclusion

What is finance?

- Alice just graduated. She has a business idea that costs 100 K€. She has no personal wealth.
- Bob just retired. He has 100 K€ of savings.
- Without finance: Bob keeps his savings under his mattress. Alice does not launch her project.
- Implications:
  - Good ideas are not implemented
  - People with savings earn no return on their savings
  - Disconnect between ideas and resources $\Rightarrow$ Poor allocation of resources

What is finance?

- With finance: Bob lends 100 K€ to Alice. Alice starts her project.
- Implications:
  - Resources are allocated to good ideas
  - This makes everyone better off
  - Countries with a developed financial sector are more prosperous (caveat: correlation $\neq$ causality)



Real-world financial arrangements

- Debt financing: Bob lends to Alice (fixed repayment + interests)

  Equity financing: Bob takes a stake in Alice's business (dividends)

- Bob may invest directly in Alice's business or through a financial intermediary (bank, fund)

- Examples

	Debt	Equity
Direct	Friends and family	Angel investors
	Bond market	Stock market
Intermediated	Bank	Equity mutual fund
	Debt mutual fund	Venture capital
		Private equity

Fundamental problem of finance

- Investors must assess if business ideas are good
- If Bob lends to Alice and Alice's project is worthless, resources are wasted (would have better been kept under the mattress or lent to someone else)

Fundamental problem of finance

A Decade After the Global Financial Crisis, Spanish Ghost Towns Remain

An estimated 3.4 million homes are currently unoccupied in Spain thanks to the country's great housing bust.

Fundamental problem of finance

- Investors must assess if business ideas are good

$=$ A prediction problem

Finance = a prediction problem

Finance = a prediction problem

Finance = a prediction problem

Use data to do prediction

Finance = a prediction problem

Example 1

- Credit score: use data to assess loan applicants' creditworthiness

experian

Social Media Insight™

Making business decisions with limited data is a huge risk. When it comes to new or emerging businesses with thin or no credit profiles, lenders must be armed with the right data to confidently and quickly assess a business.

As a breakthrough, alternative data source for risk assessment, Social Media Insight leverages social data to help lenders build a more complete picture of businesses with thin credit files.

For businesses that have thin credit profiles, a strong social media reputation can be a good measure of health. Social Media Insight aggregates social data that is directly sourced and turns it into predictive attributes:

- Number of reviews revealing if the volume trend indicates sales growth or decline.
- Unique business profile information, such as licensing, hours of operations, pricing levels, and more.
- In depth business description that goes beyond standard SIC or NAICs codes, providing potentially critical information on the type of goods or services supplied.

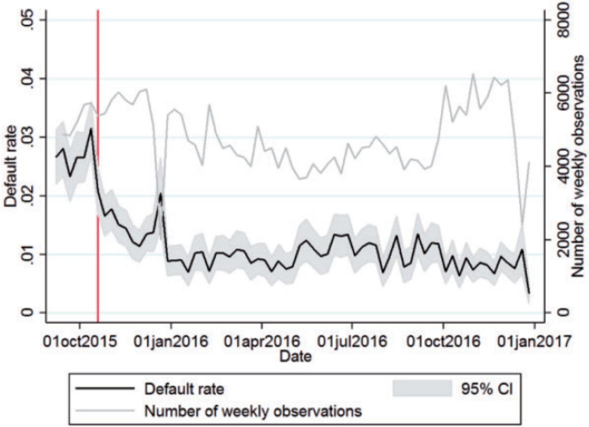
Finance = a prediction problem
Example 2

- "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 $\Rightarrow$ need to assess buyer's creditworthiness
  - Initially: credit score based on traditional information (credit history, sociodemographics)
  - After October 2015: use 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*** **	



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

Road map

What is finance?

Lending game

[optional] AI and discrimination

Conclusion

Lending game

- You run a fintech that makes loans to individuals
  - Principal amount: **10,000** paid at maturity
  - Interest rate: **r** paid upfront
  - Borrower may default (=not repay the principal) with some probability

- Your cash flow is

now	maturity
$-10,000 + 10,000 \times r$	$10,000$ if borrower repays 0 if borrower defaults

$\Rightarrow$ Your expected profit if the probability of default is p:

$$-10,000 + 10,000 \times r + (1-p) \times 10,000 + p \times 0 = 10,000 \times (r - p)$$

Example

- Default probability: 12%

  Interest rate: 8%

  Expected profit: $(0.08 - 0.12) \times 10,000 =$ loss of 400 €

- Default probability: 12%

  Interest rate: 14%

  Expected profit: $(0.14 - 0.12) \times 10,000 =$ gain of 200 €

Loan offers

- You receive 100,000 loan applications
- You decide the interest rate r to offer to each loan applicant
- You don't know the probability of default of each loan applicant $\Rightarrow$ must estimate it from data (more on this later)
- You are in competition with two other lenders (=two other players), who also make loan offers to the same pool of applicants
- Loan applicants prefer a lower interest rate but have an intrinsic preference for one of the lender. Formally:
  - Denote the three lenders' offers to applicant i by r_{i1}, r_{i2}, r_{i3}
  - 1/3 of applicants have a preference for lender 1 and choose the cheapest among $r_{i1} - 0.02$, r_{i2}, r_{i3}
  - 1/3 of applicants have a preference for lender 2 and choose the cheapest among r_{i1}, $r_{i2} - 0.02$, r_{i3}
  - 1/3 of applicants have a preference for lender 3 and choose the cheapest among r_{i1}, r_{i2}, $r_{i3} - 0.02$
  - Lenders don't know the preference of each applicant

Example

- Lender 1 offers 10%

  Lender 2 offers 11.5%

  Lender 3 offers 13%

- 1/3 of applicants (with preference for lender 1) choose lender 1

  1/3 of applicants (with preference for lender 2) choose lender 2

  1/3 of applicants (with preference for lender 3) choose lender 1

Example

- Lender 1 offers 10%

  Lender 2 offers 11.5%

  Lender 3 offers 13%

- 1/3 of applicants (with preference for lender 1) choose lender 1

  1/3 of applicants (with preference for lender 2) choose lender 2

  1/3 of applicants (with preference for lender 3) choose lender 1

- If the applicants' default probability is 11%, expected profit per offer is

  Lender 1: $\frac{2}{3} \times (0.10 - 0.11) \times 10,000 =$ loss of 66.67 €

  Lender 2: $\frac{1}{3} \times (0.115 - 0.11) \times 10,000 =$ gain of 16.67 €

  Lender 3: no gain no loss

Lending game

- The goal is to maximize profit

$$\sum_{i=1}^{100{,}000} 1(\text{Applicant } i \text{ takes your offer}) \times (r_i - 1(i \text{ defaults})) \times 10,000$$

- The key is to estimate the default probability accurately and set the interest rate accordingly

Default prediction

- NewApplications_LenderX.csv contains the 100,000 loan applications
- Lenders have **partially** overlapping information to predict default
- All three lenders have data
  - id: loan application identifier
  - sex: 1=male, 0=female
  - marital: 1=married, 0=other
  - employment: employment status (four categories)
  - income: annual income in euro (top coded at 1M euros)
- Lender $X = 1, 2, 3$ has data
  - digitalX: digital footprint score from 0 to 1 as measured by lender X's proprietary algorithm

Default prediction

- PastLoans.csv contains data on past loans with the following information
    - All the above variables, including digital1, digital2 and digital3 for all three lenders
    - default: 1=the borrower defaulted on the loan, 0=the loan was repaid
- You should use this data set to train a default prediction model

Recap of data sets

	Lender 1		Lender 2		Lender 3	
	PastLoans.csv	NewApplications.csv	PastLoans.csv	NewApplications.csv	PastLoans.csv	NewApplications.csv
sex, marital, employment, income	✓	✓	✓	✓	✓	✓
digital1	✓	✓	✓		✓	
digital2	✓		✓	✓	✓	
digital3	✓		✓		✓	✓
default	✓		✓		✓	

Offers

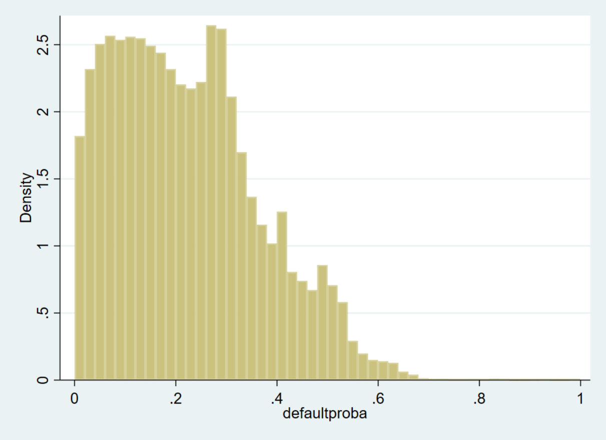
- After you have estimated a default prediction model, you make loan offers to the 100,000 applicants
- Fill in the interest rate in TeamN.csv
- The maximum allowed interest rate is 100%
- Input the interest rate as a decimal number: 0.12 for an interest rate of 12%
- You can choose not to make offers to some loan applicants. In this case, leave the cell empty
- Rename the csv file TeamN.csv with N=your team number and email it to hombert@hec.fr
- Objective: maximize profit

A scoring model

	Logit model
	Dependent variable: 1(default)
log(income)	-1.2***
	(.017)
1(income=0)	1.8***
	(.031)
student	-.74***
	(.035)
retired	-.98***
	(.03)
unemployed	-.27***
	(.03)
marital	-.15***
	(.019)
digital1	-1.3***
	(.035)
1(digital1=0)	-.61***
	(.026)
Pseudo R2	.12
N	100,000

A scoring model

- Distribution of estimated default profitability



- How to set the interest rate?

How to set the interest rate?

- You should set the interest rate <u>above</u> the estimated probability of default because of the **winner's curse**
  - Each lender has different information $\Rightarrow$ different estimate of default proba: p_1, p_2, p_3
  - If prediction models are unbiased, estimates are centered around the true default proba $\Rightarrow$ the true proba will typically lie between $\min\{p_1, p_2, p_3\}$ and $\max\{p_1, p_2, p_3\}$
  - Suppose each lender offers interest rate $=$ own estimate
  - The winning offer is the lowest one, which is $<$ true default proba $\Rightarrow$ winner makes a loss

How to avoid the winner's curve?

- Apply a margin of safety: interest rate $>$ estimated default proba

- In practice, the margin of safety can be calibrated by experimentation and back testing

Winner's curse

Example

ZDNet

Zillow: Machine learning and data disrupt real estate

Learn how big data and the Zillow Zestimate changed and disrupted real estate. It's an important case study on the power of machine learning models and digital innovation.

Written by **Michael Krigsman**, Contributor
Posted in Beyond IT Failure on July 30, 2017

Interview with Zillow's Chief Analytics Officer Stan Humphries in 2017

ZD: How accurate is the Zestimate?

S.H.: Our models are trained such that half of the Earth will be positive and half will be negative; meaning that on any given day, half of [all] homes are going to transact above the Zestimate value and half are going to transact below.

Winner's curse

Example

ZDNet

Zillow: Machine learning and data disrupt real estate

Learn how big data and the Zillow Zestimate changed and disrupted real estate. It's an important case study on the power of machine learning models and digital innovation.

Written by **Michael Krigsman**, Contributor
Posted in Beyond IT Failure on July 30, 2017

Interview with Zillow's Chief Analytics Officer Stan Humphries in 2017

ZD: How accurate is the Zestimate?

S.H.: Our models are trained such that half of the Earth will be positive and half will be negative; meaning that on any given day, half of [all] homes are going to transact above the Zestimate value and half are going to transact below.

CHRIS STOKEL-WALKER BUSINESS NOV 11, 2021 :00 AM

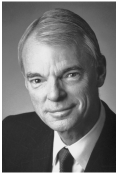
WIRED

Why Zillow Couldn't Make Algorithmic House Pricing Work

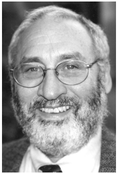
A few Nobel prizes

- The winner's curse is also called adverse selection or the lemon's problem
- It has been the subject of the 2001 and 2020 Nobel prizes

George A. Akerlof

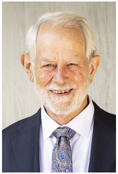


A. Michael Spence



Joseph E. Stiglitz

Paul R. Milgrom



Robert B. Wilson

Road map

What is finance?

Lending game

[optional] AI and discrimination

Conclusion

AI and discrimination

FINANCIAL TIMES

UK regulators warn banks on use of AI in loan applications

High street lenders must ensure machine learning does not worsen discrimination against minorities

Banks believe using machine learning techniques to make lending decisions could reduce discrimination against ethnic groups wh
have historically struggled to access reasonably priced loans © Chris Ratcliffe/Bloomberg

Laura Noonan in London FEBRUARY 13 2022 64

Discrimination

- The law distinguishes between direct discrimination and indirect discrimination

Direct discrimination

- Decision is based on a "protected characteristic" such as race, sex, ethnic or social origin, religion
- 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.: Lower car insurance premium for women because they have fewer accidents

- Illegal in both cases in EU and US

Indirect discrimination

- Decision is not based on protected characteristics but ends up being different for people with a protected characteristic

- Happens when the 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 social or ethnic origins

- May be legal or illegal (e.g. legal for "business necessity" in US law)

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

   Ex.: Digital footprints correlated with protected characteristics. iPhone users are 20% more likely to be women, 30% more likely to live in a city, 60% less likely to be Black (US data)

- Solution: algorithm interpretability (understand how algos make decisions)

Discriminatory algorithms?
Example

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

- Mortgage loans in the US, 2009-2015

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

  - from traditional banks: +8 basis points per year

  - from fintechs: +5 basis points per year

Discriminatory algorithms?
Example

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

Road map

What is finance?

Lending game

[optional] AI and discrimination

Conclusion

Conclusion

- Huge demand for data scientists in the financial industry

- Dual expertise in data science + finance is highly valued

- Finance @HEC
  - Master International Finance (HEC)
  - Master Data Science for Business (X-HEC)
  - Master in Economics (X-HEC-ENSAE)
  - PhD in Finance (HEC)