Hi! PARIS Summer School 2023

Tutorial 2A

Data in Finance: FinTech Lending

Johan Hombert (HEC Paris)

July 4, 2023, 8:30-12:00

Slides and data @ https://johanhombert.github.io/fintech

Road map

What is finance?

Fintech lending: Business simulation

What is finance?

- Alice just graduated. She has a business idea with setup cost 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 start her business

- Implications:

  1. Good ideas are not implemented

  2. People with savings earn no return on their savings

  $\Rightarrow$ Inefficient allocation of resources

What is finance?

- With finance: Bob lends 100 k€ to Alice, who can launch her company
- Implications:
  1. Resources are allocated to good ideas
  2. Everyone (lenders and borrower) is better off
  3. Finance key input to economic development

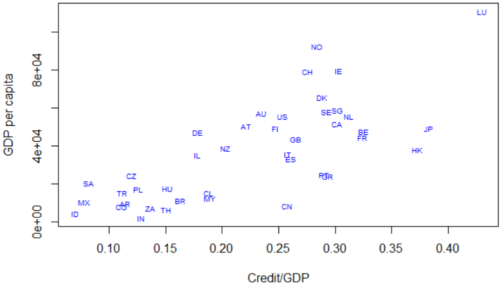


Fig.: Countries with more developed credits market are richer (caveat: correlation vs. causality)

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

- Examples

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

The fundamental problem of finance

- Investors must assess if business ideas are good

- If Bob lends to Alice and Alice's project is worthless, then resources are wasted: they would have better been kept under the mattress or lent to someone else

The 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.

The fundamental problem of finance

- Investors must assess if business ideas are good

$=$ A prediction problem

How do financiers do prediction in practice?

How do financiers do prediction in practice?

How do financiers do prediction in practice?

Use data to do prediction

Prediction with alternative data

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.

Prediction with alternative data

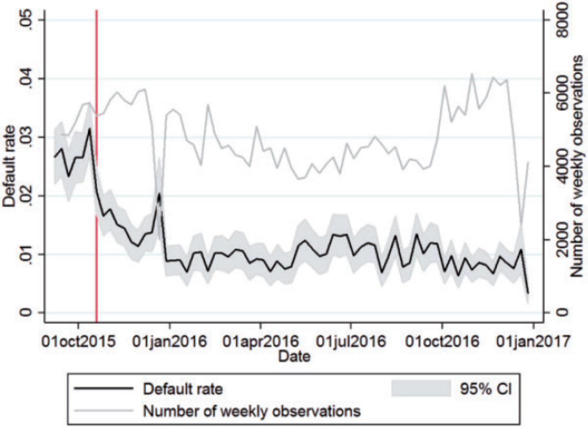
Example 2

- "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)

Variables	(1) Credit bureau bureau score Coef.	z-stat	(2) Digital footprint Coef.	z-stat	(3) Credit bureau score & digital footprint 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*** **	



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

Road map

What is finance?

Fintech lending: Business simulation

Business simulation

- You run a fintech that gives loans to individuals
  - Principal amount: 10,000 euros given to borrower now, repaid by borrower in one year
  - Interest rate: i (in %) paid upfront
  - Default risk: the borrower defaults (=does not repay the principal) with some probability p
- Your cash flow is

$\Rightarrow$ Your expected profit for a given default probability:

$$-10000 + 10000\,i + (1-p)10000 + p.0 = (i-p)10000$$

Example

- Default probability: 10%

  Interest rate: 6%

  Expected profit: $(0.06 - 0.10) \times 10,000 =$ loss of 400 €

- Default probability: 10%

  Interest rate: 12%

  Expected profit: $(0.12 - 0.10) \times 10,000 =$ gain of 200 €

Loan offers

- You receive 100,000 loan applications
- Each loan applicant has a different default probability, which you don't know $\Rightarrow$ You must estimate it from data (more on this later)
- You decide the interest rate i you offer to each loan applicant
- You are in competition with two other lenders (=two other teams), 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:
  - Let i_{k1}, i_{k2}, i_{k3} be the offers to applicant k from the three lenders
  - 1/3 of applicants have a preference for lender 1 and choose the cheapest among $i_{k1} - 0.02$, i_{k2}, i_{k3}
  - 1/3 of applicants have a preference for lender 2 and choose the cheapest among i_{k1}, $i_{k2} - 0.02$, i_{k3}
  - 1/3 of applicants have a preference for lender 3 and choose the cheapest among i_{k1}, i_{k2}, $i_{k3} - 0.02$
  - 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 (those with a preference for lender 1) choose lender 1

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

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

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

  1/3 of applicants (those with a 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

Profit

- The goal is to maximize profit

$$\sum_{k=1}^{100,000} 1\{\text{Applicant } k \text{ takes your offer}\} \times \left(i_k - 1\{k \text{ defaults}\}\right) \times 10000$$

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

Data

- 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

Data

- 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 probability 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

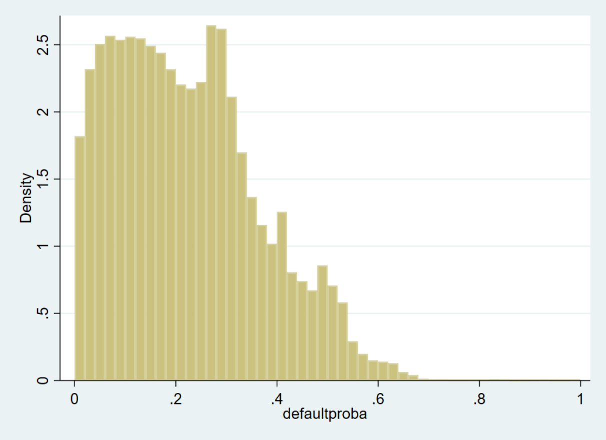
- After you have estimated a default prediction model, you make loan offers to the 100,000 applicants
- Objective: maximize profit
- The maximum allowed interest rate is 100%
- You can choose not to make offers to some loan applicants
- Input the offers in a csv file with two columns
  - id: from the original dataset, from 1 to 100,000
  - rate: your interest rate inputted as a decimal number (0.12 for an interest rate of 12%). Leave the cell empty if you do not make an offer to a given loan applicant
- Name the csv file TeamN.csv where N is your team number (between 11 and 35) and email it to hombert@hec.fr

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 ABOVE 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 below the true default proba $\Rightarrow$ The winner always makes a loss!

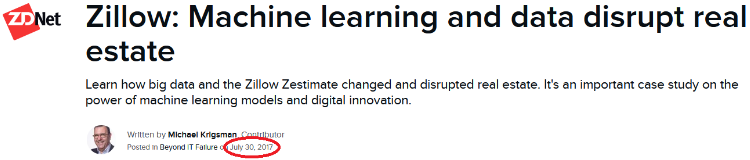
How to avoid the winner's curve?

- 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

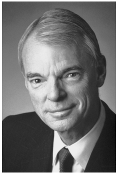
WIRED

Why Zillow Couldn't Make Algorithmic House Pricing Work

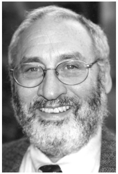
A few Nobel prizes

- 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

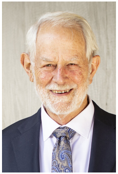


A. Michael Spence



Joseph E. Stiglitz

Paul R. Milgrom



Robert B. Wilson