Tackling the Challenges of Big Data
Big Data Analytics

Andrew W. Lo
Charles E. and Susan T. Harris Professor
Massachusetts Institute of Technology

Tackling the Challenges of Big Data
Big Data Analytics
Applications: Finance
The Challenge of Consumer Credit Risk Management

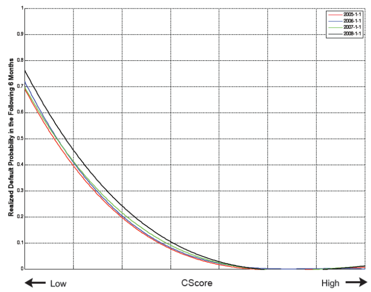
Andrew W. Lo
Charles E. and Susan T. Harris Professor
Massachusetts Institute of Technology

Consumer Credit Risk Management

- $3T of consumer credit outstanding as of 8/13
- $840B of it is revolving consumer credit
- Average credit card debt as of 10/13: $15,159
- 46.7% of households carry positive credit card balance as of 12/12
- Current "charge-off" rates are 6.7% (2013Q2), but reached 10.2% in 2010Q1

⇒ **Can We Predict These Credit Cycles?**

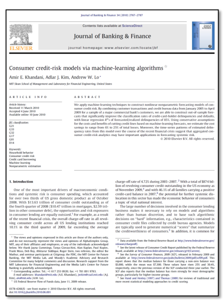
Standard Credit Scores Are Too Insensitive



Tackling the Challenges of Big Data

MIT Laboratory for Financial Engineering:

"Consumer Credit Risk Models via Machine-Learning Algorithms", by
Amir E. Khandani,
Adlar J. Kim,
Andrew W. Lo
Journal of Banking & Finance (2010)



Tackling the Challenges of Big Data

Tackling the Challenges of Big Data

Big Data Analytics

Applications: Finance

The Challenge of Consumer Credit Risk Management

THANK YOU

PROFESSIONAL EDUCATION

CSAIL

Tackling the Challenges of Big Data

Big Data Analytics

Andrew W. Lo
Charles E. and Susan T. Harris Professor
Massachusetts Institute of Technology

PROFESSIONAL EDUCATION
CSAIL

Tackling the Challenges of Big Data

Big Data Analytics

Applications: Finance

Big Data for Consumer Credit

Andrew W. Lo
Charles E. and Susan T. Harris Professor
Massachusetts Institute of Technology

PROFESSIONAL EDUCATION
CSAIL

Anonymized Data from Large U.S. Commercial Bank

Transaction Data

	By Category		
Transaction Count	Mortgage payment	Hotel expenses	Bar Expenses
Total Inflow	Credit car payment	Travel expenses	Fast Food Expenses
Total Outflow	Auto loan payment	Recreation (golf	Total Rest/Bars/Fast-Food
	Student loan payment	Department Stores Expenses	Healthcare related expenses
By Channel:	All other types of loan payment	Retail Stores Expenses	Health insurance
	Other line of credit payments	Clothing expenses	Gas stations expenses
ACH (Count, Inflow and Outflow)	Brokerage net flow	Discount Store Expenses	Vehicle expenses
ATM (Count, Inflow and Outflow)	Dividends net flow	Big Box Store Expenses	Car and other insurance
BPY (Count, Inflow and Outflow)	Utilities Payments	Education Expenses	Drug stores expenses
CC (Count, Inflow and Outflow)	TV	Total Food Expenses	Government
DC (Count, Inflow and Outflow)	Phone	Grocery Expenses	Treasury (eg. tax refunds)
INT (Count, Inflow and Outflow)	Internet	Restaurant Expenses	Pension Inflow
WIR (Count, Inflow and Outflow)	Collection Agencies	Unemployment Inflow	Collection Agencies

Balance Data

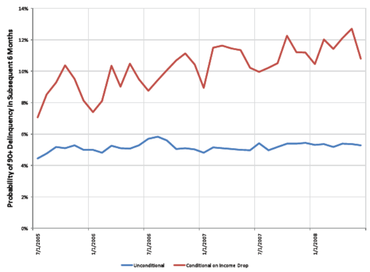
Checking Account Balance
Brokerage Account Balance
Saving Account Balance
CD Account Balance
IRA Account Balance

Credit Bureau Data

File Age	Type (CC, MTG, AUT, etc)
Credit Score	Age of Account
Open/Closed Flag & Date of Closure	Balance
Bankruptcy (Date & Code)	Limit if applicable
MSA & Zip	Payment Status
	48-Month Payment Status History

1% Sample = 10 Tb!

Extract "Interesting" Features



Tackling the Challenges of Big Data

Tackling the Challenges of Big Data
Big Data Analytics
Applications: Finance
Big Data for Consumer Credit

THANK YOU

Tackling the Challenges of Big Data
Big Data Analytics

Andrew W. Lo
Charles E. and Susan T. Harris Professor
Massachusetts Institute of Technology

Tackling the Challenges of Big Data

Big Data Analytics

Applications: Finance

Machine Learning Techniques for Analyzing Big Data

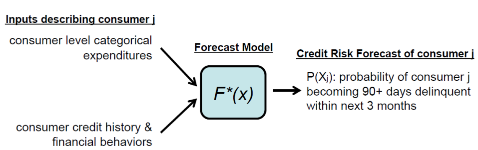
Andrew W. Lo

Charles E. and Susan T. Harris Professor

Massachusetts Institute of Technology

Objectives

- For consumer j with characteristics or "features" X_j, estimate probability of default or delinquency $P(X_j)$
- Characteristics include:
  - Individual characteristics, macro factors, interactions between the two

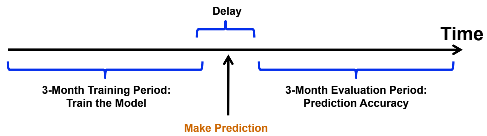


Machine Learning Techniques

- Decision trees (e.g., CART)
- Logistic regression
- Random forests
- Clustering/segmentation (can be used with other models)
- Software:
  - WEKA (machine-learning suite – University of Waikato, NZ) http://www.cs.waikato.ac.nz/ml/weka/
  - LIBLINEAR (National Taiwan University) http://www.csie.ntu.edu.tw/~cjlin/liblinear/

Model Evaluation Framework

- Prediction made for probability of going 90+ delinquent for individual credit cards over 3-month horizon
- Using non-overlapping data (in time) to calibrate the model:



Model Evaluation Framework

Training Period			Evaluation Period		
	Delinquency Period		Prediction	Delinquency Period	
Input Date	Start Date	End Date	Date	Start Date	End Date
Jan-08	Feb-08	Apr-08	Apr-08	May-08	Jul-08
Feb-08	Mar-08	May-08	May-08	Jun-08	Aug-08
Mar-08	Apr-08	Jun-08	Jun-08	Jul-08	Sep-08
Apr-08	May-08	Jul-08	Jul-08	Aug-08	Oct-08
May-08	Jun-08	Aug-08	Aug-08	Sep-08	Nov-08
Jun-08	Jul-08	Sep-08	Sep-08	Oct-08	Dec-08
Jul-08	Aug-08	Oct-08	Oct-08	Nov-08	Jan-09
Aug-08	Sep-08	Nov-08	Nov-08	Dec-08	Feb-09
Sep-08	Oct-08	Dec-08	Dec-08	Jan-09	Mar-09
Oct-08	Nov-08	Jan-09	Jan-09	Feb-09	Apr-09

Summary Statistics

Starting Date	Ending Date	Total Credit Cards Count	Customers Going 90+ Days Delinquent		Customers **NOT** Going 90+ Days Delinquent	
			Count	Percent of Total	Count	Percent of Total
May-08	**Jul-08**	575,573	13,939	2.4	561,634	97.6
Jun-08	**Aug-08**	644,396	14,112	2.2	630,284	97.8
Jul-08	**Sep-08**	680,928	14,087	2.1	666,841	97.9
Aug-08	**Oct-08**	720,285	14,371	2.0	705,914	98.0
Sep-08	**Nov-08**	720,660	14,880	2.1	705,780	97.9
Oct-08	**Dec-08**	718,465	14,971	2.1	703,494	97.9
Nov-08	**Jan-09**	715,943	15,202	2.1	700,741	97.9
Dec-08	**Feb-09**	710,732	16,579	2.3	694,153	97.7
Jan-09	**Mar-09**	661,006	16,984	2.6	644,022	97.4
Feb-09	**Apr-09**	659,342	16,689	2.5	642,653	97.5

Tackling the Challenges of Big Data
Big Data Analytics
Applications: Finance
Machine Learning Techniques for Analyzing Big Data

THANK YOU

MIT PROFESSIONAL EDUCATION CSAIL

Tackling the Challenges of Big Data
Big Data Analytics

Andrew W. Lo
Charles E. and Susan T. Harris Professor
Massachusetts Institute of Technology

MIT PROFESSIONAL EDUCATION CSAIL

Tackling the Challenges of Big Data
Big Data Analytics
Applications: Finance
Empirical Results for a
Commercial Bank's Credit Card Division

Andrew W. Lo
Charles E. and Susan T. Harris Professor
Massachusetts Institute of Technology

MIT PROFESSIONAL EDUCATION CSAIL

Empirical Results

Training Window		Prediction Date	Evaluation Window		Average Predicted Probability of Going 90+ Delinquent		
Start	End		Start	End	Among All	Among Customers Going 90+ Days Delinquent	Among Customers **NOT** Going 90+ Days Delinquent
Feb-08	Apr-08	**Apr-08**	May-08	Jul-08	2.5	**61.2**	1.0
Mar-08	May-08	**May-08**	Jun-08	Aug-08	2.0	**62.1**	0.6
Apr-08	Jun-08	**Jun-08**	Jul-08	Sep-08	1.9	**60.4**	0.7
May-08	Jul-08	**Jul-08**	Aug-08	Oct-08	1.9	**62.5**	0.6
Jun-08	Aug-08	**Aug-08**	Sep-08	Nov-08	2.0	**62.4**	0.7
Jul-08	Sep-08	**Sep-08**	Oct-08	Dec-08	2.1	**63.6**	0.8
Aug-08	Oct-08	**Oct-08**	Nov-08	Jan-09	2.1	**62.5**	0.8
Sep-08	Nov-08	**Nov-08**	Dec-08	Feb-09	2.2	**59.8**	0.8
Oct-08	Dec-08	**Dec-08**	Jan-09	Mar-09	2.4	**60.8**	0.8
Nov-08	Jan-09	**Jan-09**	Feb-09	Apr-09	2.4	**62.8**	0.9

Captures overall increase in risk

Shows clear separation between two types of customers

Empirical Results

- Type I and Type II error tradeoffs can be controlled by varying the threshold of the model.

Classifier Threshold = 10%

Actual Outcome	Model Prediction: Good	Model Prediction: Bad
Good	95.16%	2.27%
Bad	0.32%	2.25%

Classifier Threshold = 20%

Actual Outcome	Model Prediction: Good	Model Prediction: Bad
Good	96.37%	1.06%
Bad	0.44%	2.13%

Classifier Threshold = 30%

Actual Outcome	Model Prediction: Good	Model Prediction: Bad
Good	96.78%	0.65%
Bad	0.57%	2.00%

Classifier Threshold = 50%

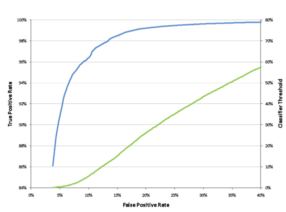
Actual Outcome	Model Prediction: Good	Model Prediction: Bad
Good	97.14%	0.29%
Bad	0.89%	1.68%

Empirical Results

- **Receiver Operating Characteristic** (ROC) Curve
- Summarizes the trade-off noted on last slide
- True and false positive rate is calculated for different level of threshold

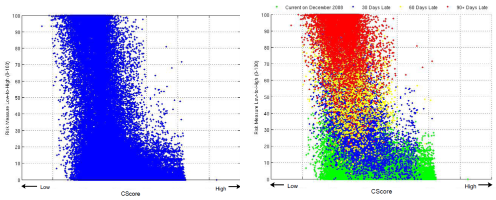
The threshold level can be optimized based on:

- Business objectives
- Risk appetite
- Capital requirements
- Employment cycle
- Etc.



Empirical Results

- Comparison with traditional credit scores



Tackling the Challenges of Big Data

Tackling the Challenges of Big Data

Big Data Analytics

Applications: Finance

Empirical Results for a
Commercial Bank's Credit Card Division

THANK YOU

Tackling the Challenges of Big Data

Big Data Analytics

Andrew W. Lo
Charles E. and Susan T. Harris Professor
Massachusetts Institute of Technology

Tackling the Challenges of Big Data

Big Data Analytics

Applications: Finance

Gauging the Practical Value of Big Data for Consumer Credit Risk Management

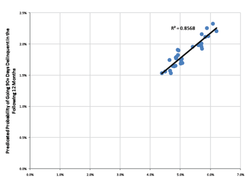
Andrew W. Lo

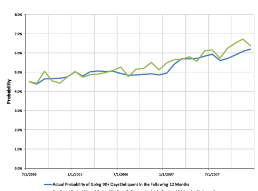
Charles E. and Susan T. Harris Professor

Massachusetts Institute of Technology

Macro Forecasts of Credit Losses

- Forecasts of future credit losses may be used to construct an early warning system (12 months ahead!) for emerging problems in consumer credit





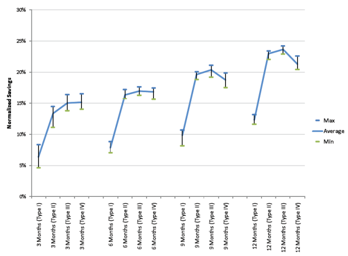
Measuring Value-Added of Forecasts

Assume that:

- In the beginning, both good and bad consumers will have the same average running balance
- Bad consumers will incur certain rate of "run-up" in their balance before default (we use 10%, 20%, 30% and 50% in our analysis)
- Credit card interest rate and lender's funding cost rate are fixed at 5%
- Time horizon over which consumers amortize their credit card balance is fixed (we use 3, 5, and 10 years)
- **The estimated value-added ranges from 6% to 24%, depending on the assumed parameters and client type (see next slide)**

Measuring Value-Added of Forecasts

- Type I clients have "thin" files (very few transactions), Type IV clients have very "thick" files (many transactions)



These results show that the availability of features makes a big difference in forecast power and value-added

Tackling the Challenges of Big Data

Conclusion

- Big data can be used to construct better consumer credit risk forecasts
- Machine-learning techniques can add value
- High dimensionality of the data is both a blessing and a curse
- Key aspects are feature-vector construction and nonlinear interactions
- Many practical applications are possible

Tackling the Challenges of Big Data

Tackling the Challenges of Big Data

Big Data Analytics

Applications: Finance

Gauging the Practical Value of Big Data for Consumer Credit Risk Management

THANK YOU

Tackling the Challenges of Big Data
Big Data Analytics

Andrew W. Lo
Charles E. and Susan T. Harris Professor
Massachusetts Institute of Technology

MIT PROFESSIONAL EDUCATION

CSAIL