Mobile Phone-based Credit Scoring

Skyler Speakman
Research Scientist
Mobile Phone-based Credit Scoring

Skyler Speakman
Research Scientist
IBM Research - Africa
**Blockchain** can address many challenges facing developing countries, e.g.:

- Decentralized school information hub (e.g., students, teachers, assets) management
- Proof of ownership and a marketplace for sales and purchase of digital assets (e.g., land title)
- Decentralized cross-boarder management
- Digitization of trade/supply documents and workflows
- Identity companion for enabling anywhere anytime trusted multi-factor identity verification
- Digitization of documents and proof of ownership for transfers
- Decentralized Tax Collections using a network of Point of Sales/Transactions devices on blockchain
EPiC: Public Health

Disease surveillance
1. Decision support
2. Descriptive analytics
3. Smart data search

Citizen engagement
4. Mobile data collection
5. On-demand language analytics
Doing Business in Kenya has been made easier by streamlining procedures, time, costs and laws associated with the Doing Business Indicators.
Inclusive Financial Services

Payments  Savings  Credit  Insurance  Investments

Complexity
Inclusive Financial Services

How can mobile phone data help?
Access to financial services such as credit, savings, insurance is a global challenge.

2-3 BILLION individuals and 200 MILLION businesses in emerging economies today lack access to savings and credit, and even those with access can pay dearly for a limited range of products.

[McKinsey Global Institute, Digital finance for all: Powering inclusive growth in emerging economies, 2016]
Credit Scoring

- Past cell phone data
- Monitoring
- Repaid all loans
- Defaulted at least one loan
Credit Scoring

- Past cell phone data
- Monitoring
- Repaid all loans
- Defaulted at least one loan

Reduce Defaults by 55%
Credit Scoring

Recommend 1M Customers

- Past cell phone data
- Monitoring

Registration Date

Time

- Repaid all loans
- Defaulted at least one loan

Reduce Defaults by 55%
Monkey in Kenya Survives After Setting Off Nationwide Blackout

A nationwide blackout that knocked out lights, severed internet service and paralyzed countless businesses for more than three hours on Tuesday was caused by a monkey, the electricity company announced.
Life in Africa
Life in Africa
Before M-Pesa

(Thanks to Dr. Tavneet Suri for this slide)
• Increases savings
• Increases consumption and reduces poverty
• Does not increase (physical) assets
• Increases use of a bank account
• Changes occupational choice (less likely to be a farmer, more likely to be in a business), especially for women in the household

(Thanks to Dr. Tavneet Suri for this slide)
8 Years After M-Pesa

(Thanks to Dr. Tavneet Suri for this slide)
An IBM banking partner in Kenya has a savings and credit product that transfers money through M-PESA.

Some customers are given a credit limit and may borrow from the bank. Funds are available in a few seconds through the M-PESA channel.

All transactions, including signing-up, are completed over the phone. No need to set foot in a traditional bank.
An IBM banking partner in Kenya has a savings and credit product that transfers money through M-PESA. Some customers are given a credit limit and may borrow from the bank. Funds are available in a few seconds through the M-PESA channel.

All transactions, including signing-up, are completed over the phone. No need to set foot in a traditional bank.
An IBM banking partner in Kenya has a savings and credit product that transfers money through M-PESA. Some customers are given a credit limit and may borrow from the bank. Funds are available in a few seconds through the M-PESA channel.

All transactions, including signing-up, are completed over the phone. No need to set foot in a traditional bank.
An IBM banking partner in Kenya has a savings and credit product that transfers money through M-PESA.

Some customers are given a credit limit and may borrow from the bank. Funds are available in a few seconds through the M-PESA channel.

All transactions, including signing up, are completed over the phone. No need to set foot in a traditional bank.
Mobile Money Withdrawal Amount in 6 Months (std)

Data
Airtime Usage

![Airtime Usage in 6 Months (std)](image)

- No. of Customers
- Airtime Usage (50 USD) (230 USD)

Data
Decision Trees

Split data by choosing a feature and cutoff that separate the two classes

Multiple splits can identify interaction effects

Many advantages:
- Classification or regression
- Real-valued or discrete predictors
- Easily handles missing data
- Scales to large data sets
- Ignores irrelevant features
Train a decision tree
(weak learner)
Reweight data to prioritize the misclassifications
Reweight data to prioritize the misclassifications

Train an additional tree on reweighted data
Reweight data to prioritize the misclassifications

Boosting Decision Trees
Reweight data to prioritize the misclassifications

Train an additional tree on reweighted data
A boosted *sequence* of weak learners becomes a strong learner

Boosting maintains advantages of decision trees while increasing classification accuracy
Boosting Parameters

How many trees?
How deep for each tree?

Cross validation is used to explore the space.

180 Trees of Depth 2 maximized AUC.
Results – Decreasing Defaults
Results – Decreasing Defaults

Identified 55% of would-be defaulters

Retained 83% of paying customers

(specificity)

(sensitivity or recall)
Results – Increasing Revenue

The graph shows the distribution of customers based on their probability of repaying, with two categories: > 0 Initial Limit and 0 Initial Limit. The x-axis represents the probability of repaying, ranging from 0.0 to 1.0, while the y-axis represents the number of customers. The graph depicts a bell-shaped curve for each category, indicating a normal distribution.
Results – Increasing Revenue

+1 Million customers allowed on credit product
Airtime Borrowed

Loans are AUTOMATICALLY repaid when topping up again

Higher probability of repayment

Lower probability of repayment

Airtime Borrowed in 6 Months (std)
• Machine Learning methods are sensitive to changes between training and testing data.
• Trained on Kenya data but rolled out on Uganda data.
  – Mitigated by standardizing variables and making decisions based on ‘how many standard deviations from the mean’
Transfer Learning

• Implications for 3 Important Areas:
  – Client: Has expanded into multiple markets with the same ‘task’ but different demographics
  – Business: Many problems span the continent but with country-specific nuances. Transfer Learning skills help scale in this environment.
  – Research: It is a difficult problem and open area of research.
Try to minimize the (cost of) mistakes between predicted class and actual class (empirical risk).

In other words, try to model the joint distribution of features X and outcome Y.

Typically assume this joint distribution is composed of two parts: How X’s are distributed, and how Y’s relate to that distribution.

What happens if the distribution of X’s change in the training and test data?

\[
\hat{Y}(\cdot) = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f(x_i), y_i)
\]

\[p_{X,Y} = p_X p_{Y|X}\]
Covariate Shift

Number of days with low airtime balance (6 months)

- Kenya
- Uganda

Class

Count

Days

0 50 100 150 200

0 2,000 4,000 6,000 8,000 10,000
Covariate shift re-weights the observations so that the training data 'looks like' the shifted test data.

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{q_X(x_i)}{p_X(x_i)} \mathcal{L}(f(x_i), y_i)
\]
The problem: We do not even have $q_x$!
The problem: We do not even have $q_x$!
(Generalized) Covariate Shift

The problem: We do not even have $q_x$!
(Generalized) Covariate Shift

\[ \hat{Y}(\cdot) = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f(x_i), y_i) \]

Assumes customers default for same reasons in both countries

Uganda Users

\[ \frac{1}{n} \sum_{i=1}^{n} \frac{q_X(x_i)}{p_X(x_i)} \mathcal{L}(f(x_i), y_i) \]

Kenya Users

Assumes customers ‘self select’ in to product for same reasons in both countries

Uganda Population

\[ \sum_{x} \hat{p}_{X|E,M}(x \mid 1, 0) \frac{\hat{p}_{X|M}(x \mid 1)}{\hat{p}_{X|M}(x \mid 0)} \frac{1}{n(x)} \sum_{i: x_i = x} \mathcal{L}(f(x), y_i) \]

Kenya Users

Kenya Population
• Extensions
  – Only implemented in least squares (regression)
    • What about boosting? Altering the initial weights of observations the same as reweighting empirical risk?
  – Only analyzed in two domains (original, new)
    • Can we extend to k-domains? Can 3 countries be used to improve accuracy on unknown 4th?
  – Only worked with discretized features
    • Continuous inputs are theoretically possible and should be explored.
  – How does it compare to online methods that learn as new data comes in (trAdaboost)?
Strong performance. No TL needed.

Should be strong but only weak results. No TL needed.

Generalized Covariate Shift. Lower expectations due to poor Uganda performance.

Live in Uganda for first 10 months. Standardizing X’s only. Poor performance.
Incorporating Temporal Data

Kenya
- 6 Months Telco Data
- Registration
- Loan Request

Uganda
- Rich panel data for feature extraction
- Monthly update of Telco data
- Up-to-date data
- Loan Request

Less relevant data & Harder to predict

Up-to-date data Uganda

6 Months Telco Data