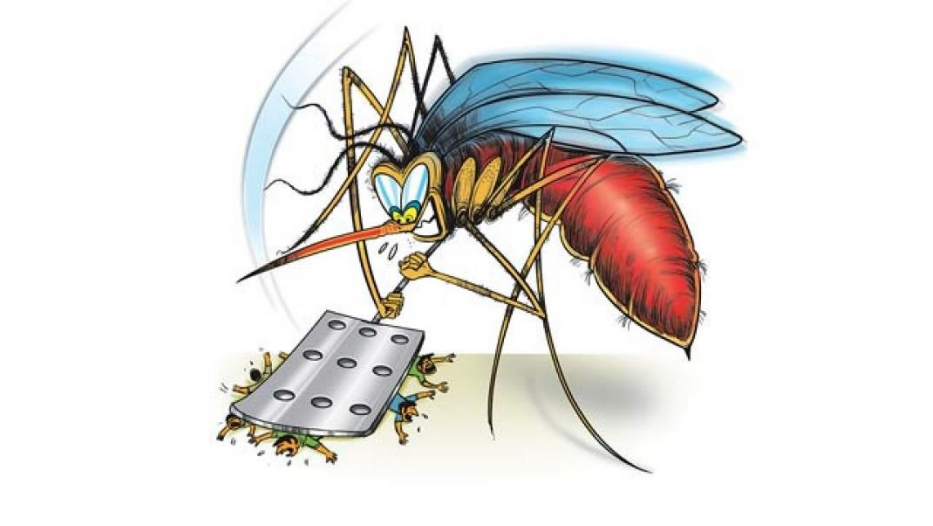
Data for Development

Lecture 26: CSE 490c



Announcements

- Homework 7, Due December 3
- Programming Assignment 4, Due December 10
- Readings for Wednesday / Friday posted

Topics

- Data Science for Development
- AI for Social Good

- Today
  - Predicting Dengue fever based on helpline calls
  - Credit scoring

Forecasting dengue fever outbreaks based on telephone hotline

RESEARCH ARTICLE

PUBLIC HEALTH

Fine-grained dengue forecasting using telephone triage services

Nabeel Abdur Rehman,[1,2] Shankar Kalyanaraman,[3,4] Talal Ahmad,[3,4] Fahad Pervaiz,[5] Umar Saif,[1,6] Lakshminarayanan Subramanian[3,4]*

- Contribution: Show that hotline calls can be used for localized prediction of dengue cases when combined with an environmental model
- Calls provide a two week leading indicator

Goal: Predict cases at a local level based on call data

- How does calls to hotline predict suspected dengue cases
  - Comparing two measured quantities
- Are other variables needed?
  - Weather information
  - Awareness campaigns for hotline
- Localize across 10 towns of Lahore

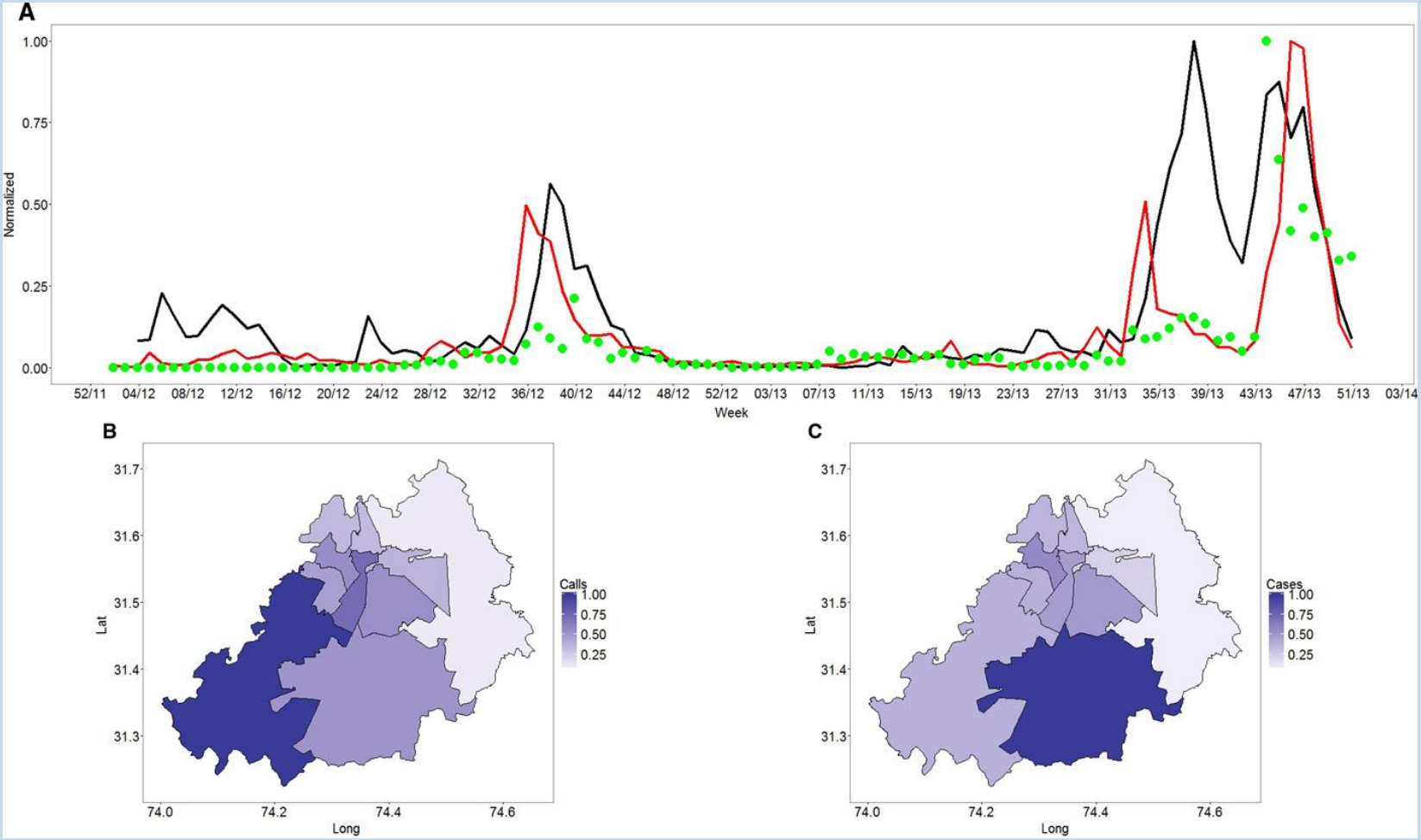


Fig. 1. Trends in call volume and suspected dengue cases measured during 2012 and 2013. (**A**) Time series of calls (red), suspected dengue cases (black), and awareness campaigns (green points). Scale normalized by dividing by individual maximum values. The *x*-axis label is in week of the year. (**B**) Density map of calls across towns in Lahore. (**C**) Density map of cases across towns in Lahore. The lightest shade represents the least number, and the darkest shade represents the highest number. The legend is normalized by the maximum value. Lat, latitude; long, longitude.

Nabeel Abdur Rehman et al. Sci Adv 2016;2:e1501215

ScienceAdvances AAAS

Variables

Variable	Meaning
S(w,t)	Suspected cases week w, town t
C(w,t)	Calls week w, town t
A(w,t)	Awareness campaigns week w, town t
H(w)	Humidity week w
T(w)	Temperature week w
R(w)	Rainfall week w

Predict log S(w + 2, t) as a function of C(w,t), A(w,t), H(w), T(w), R(w)

Methodology

- Linear regression
- Random forest learning algorithm

- Estimates achieved good fit based on root mean square error
- Most important term, number of calls

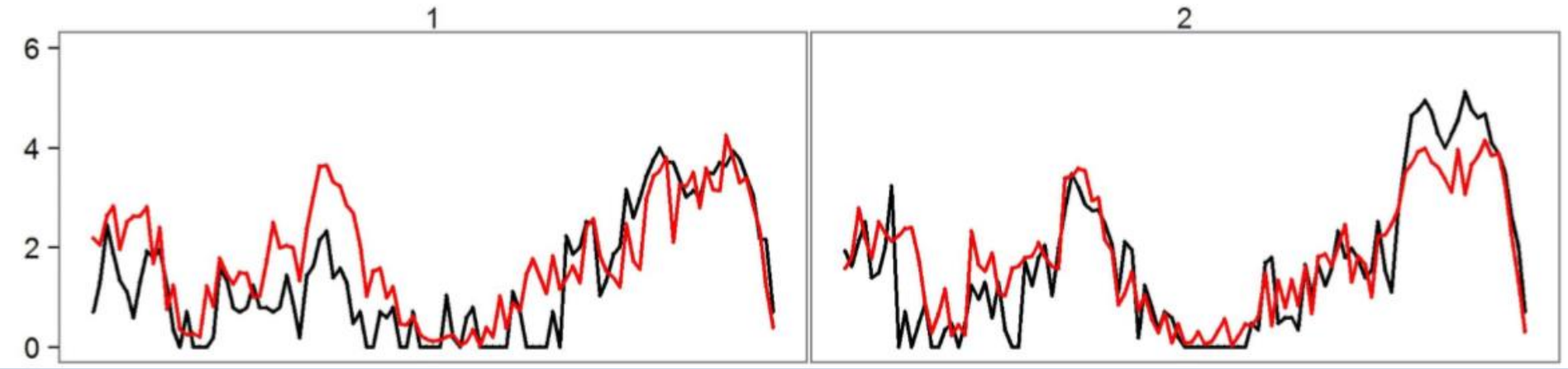


Fig. 2. Town-wise predictions of log-suspected cases from the ensemble model based on calls and weather data. Suspected dengue cases (black) and predictions from the model (red).

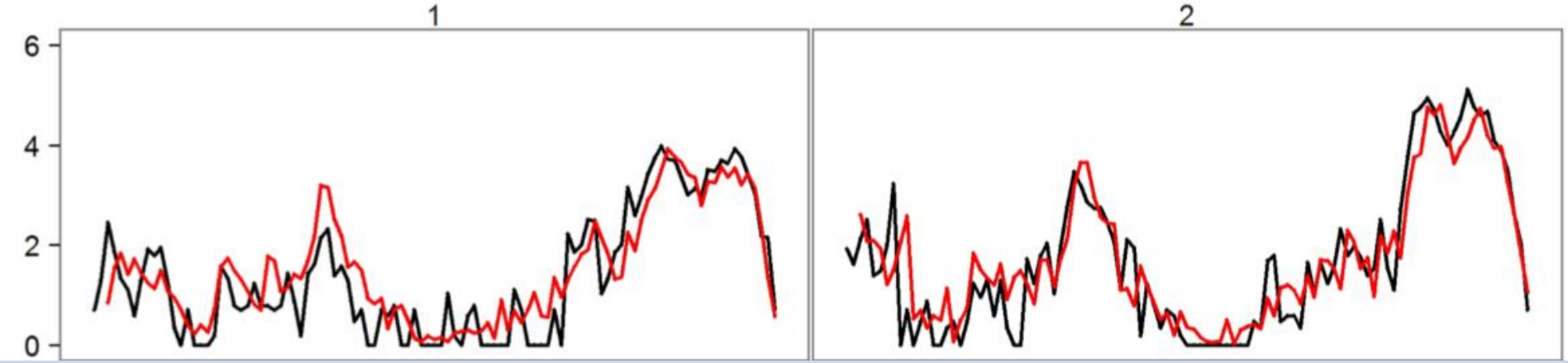
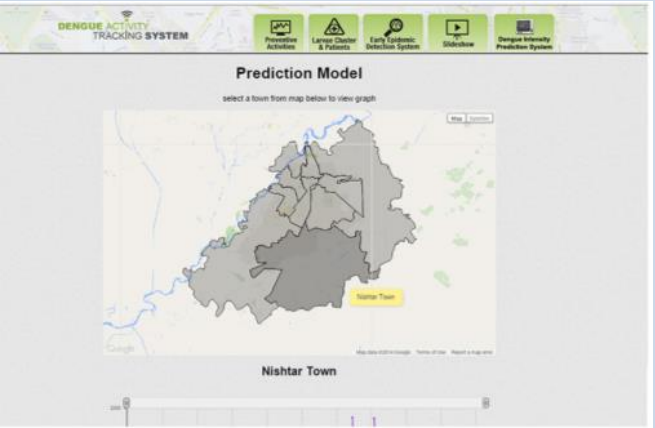


Fig. 3. Town-wise predictions of log-suspected cases from the ensemble model based on calls, cases, and weather data. Suspected dengue cases (black) and predictions from the model (red).

Validation and Deployment

- Validation done through standard methods of generating models for randomly constructed subsets and confirming on held out data
- Validation and cross-correlations as a supplement to the journal article
- Model deployed as part of dengue surveillance system, accessible through a web api



Credit scoring

- Mechanism to evaluate how likely a person is to repay a loan
- Computed by a magic algorithm based on a wide amount of financial data
- How can credit worthiness be determined in settings without strong financial infrastructure
  - Collateral
  - References

Financial Inclusion

- Payments
- Savings
- Credit
- Insurance
- Investment

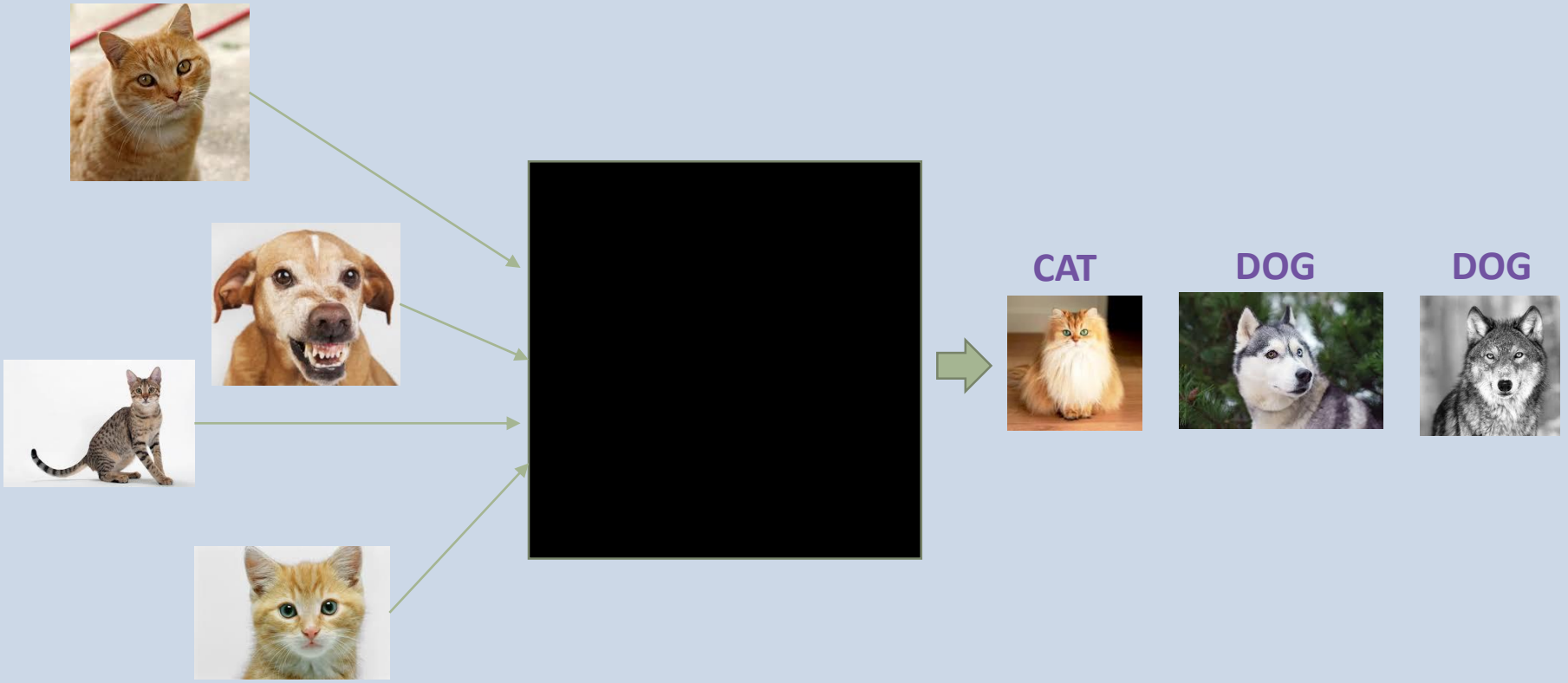
Mobile loans

- Loans associated with mobile phone or mobile money
- Easy access through App or mobile money menus
- Integrated with mobile money
- Short term, high interest

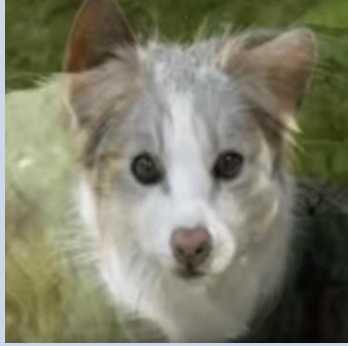
Granting credit

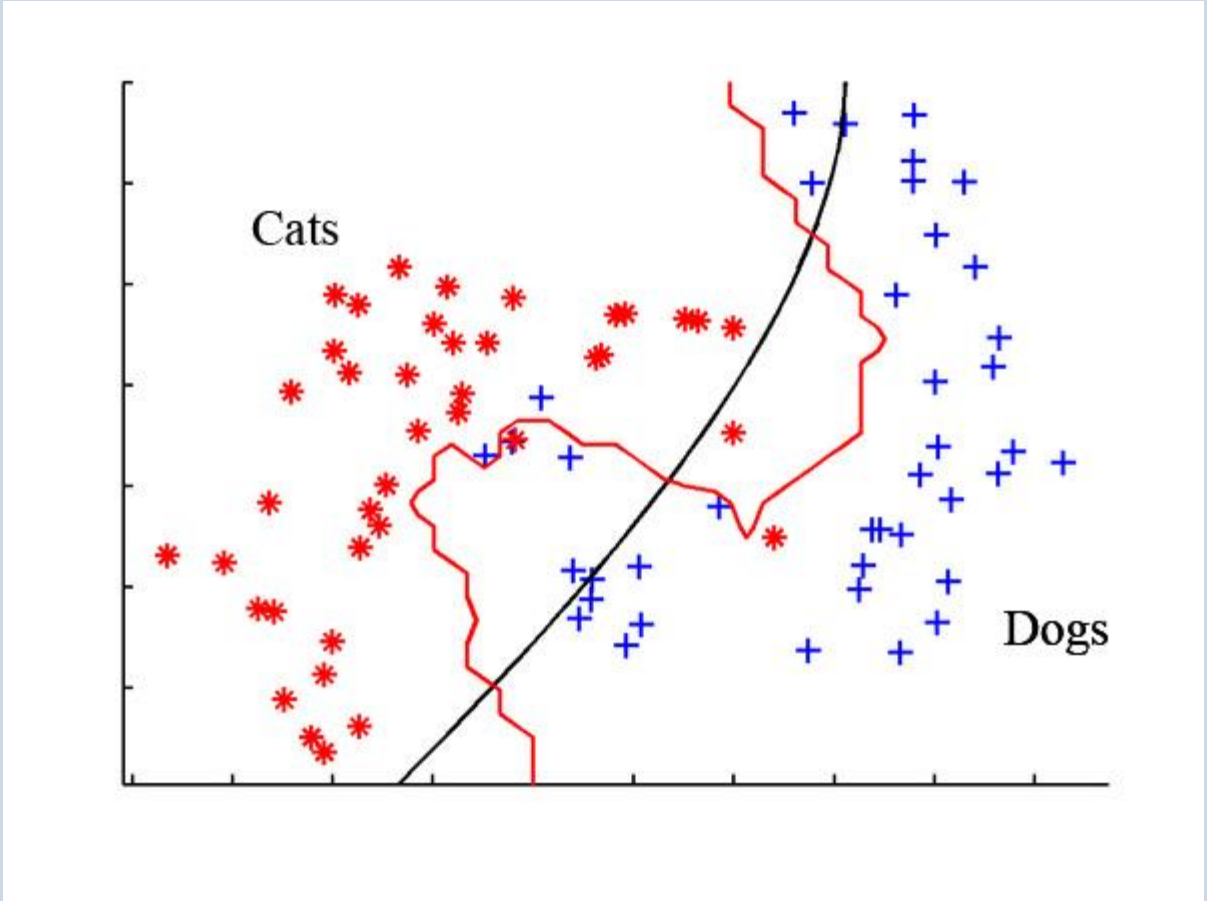
- Repayment rate
  - Grant low value loans with limited security
  - Increase loan amounts based on successful payment history
- Phone usage
  - Length of account ownership (e.g., at least six months)
  - Recharge history
- Mobile money usage
  - Number and amount of transactions
  - Many Apps access mobile money history through SMS receipts
- Smartphone usage
  - Installed applications
  - Phone model
- Blacklist

Machine Learning



Standard Modelling

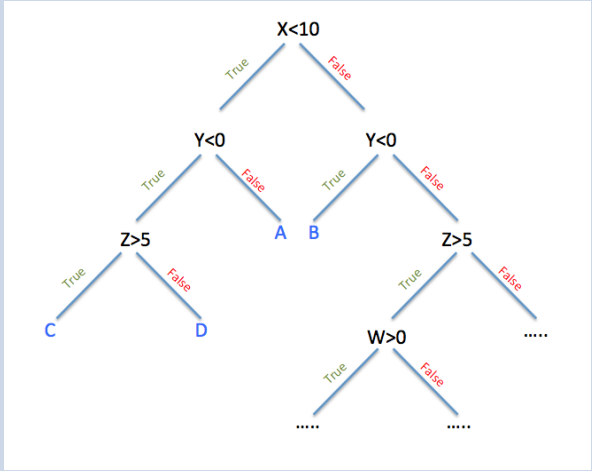




General Methodology

- Build a training set of labeled examples
- Create a model that does a good job on training data
- Use the model for new data

Types of models



- Regression
- Decision trees
- Reweighting
- Neural Networks

- Models are mathematical functions from the domain which can be of very high dimension
- Constructing the model may be computationally intensive

Features

- Modelling depends on choice of features
  - Image given as pixels
  - Image pre-processed with edge detection
  - Financial data as a collection of transactions
  - Financial data as an aggregate of transactions such as total withdrawals
- High dimensional models may derive “features” from the data set

Goals of lenders

- Making money . . .
- In theory, accurately predicting repayment performance allows setting more attractive terms to compete for customers
- Revenue optimization gives an objective function

ML and credit scoring

- Multiple papers have been written on the subject
  - Results hard to assess because real data is proprietary
- Reported correlations between length of phone ownership, number of calls and loan default
- Repayment history is important – but want to establish score before giving any loans
- Are there underlying features, such as employment status or wealth that can be predicted from cell phone data?

From IBM **Research** | Africa