Scope of Work for Data Science Final Project

Prepared by Group #12
Alice Smith, alice.smith@fas.harvard.edu
Bob Doe, robert.doe@fas.harvard.edu
Charlie Brown, charlie.brown@fas.harvard.edu

Project Statement and Background

Doogle is a micro-financial institution that makes thousands of micro-loans each month to people in the developing world. They have been making these loans for four years, and have been using roughly the same go/no-go decision flowchart for the past two years.

These loans, typically ranging from $100 to $2,000 in size and 2 months to 6 months in duration are used by the borrower to make small capital investments, for example to buy livestock or crop seeds which will then generate revenue, to receive professional training for licensure, or as capital investment for materials necessary to start a small business such as a mobile phone airtime stand.

Like many micro-loans, the loans that the company makes are unsecured, which means that the borrower provides no collateral and therefore all of the counterparty risk is borne by the company. Loan defaults, while rare, are the single most serious threat to the long term sustainability of the company's business model. Doogle believes based on industry research that their current borrower default rate of 3.2%, while not unsustainable, is high for the micro-financial industry.

Goals: Using past data on loan applications and outcomes, optimize the company?s lending decision model for whether to lend or not lend to improve the current default rate by at least 70 basis points (down to less than 2.5%). They also want a system which automatically runs new applications through the model and displays this recommendation to loan officers along with justification for decision in a user friendly way.

Literature Review

- Paper 1
- Paper 2

Available resources/data

Over 12,000 records of loan applications, will be a structured csv file containing:

- Which loan officer was assigned to the case.

- Basic personal and financial information (e.g. age, gender, geographic location, number of children, history of previous loans from Client if any, self reported cash on hand, self reported loan history, forms of government identification confirmed by the lender, etc).
- Requested loan amount (in USD).
- For a sample of loan applications (4,000), a third party credit score service was used to get an additional indicator. This is a numeric value scaled between 200 and 900.

  Our data is stored in this location: http://ourdata

  Our plan to explore and clean the data is the following:

Preliminary EDA

A simple visualization with a brief explanation of the data will do. You may use matplotlib or seaborn or Tableau.