Machine Learning for Social Good: Predicting Success of Fundraising Projects Using the DonorsChoose Data Set

Jack Graham, Andrew Spencer, Kevin Tsao, Yanjun Yang

Motivation and Goal

- Predict the success of fundraising projects through DonorsChoose
  - Online platform connecting educators looking for funding and donors willing to help
  - Determine characteristics of project requests that can help a project succeed
- Data on 1.2 million projects, of which 75% were successful
  - Contains project titles, essays, and short descriptions in paragraph form
  - Data on locations, project success, and type of project being requested

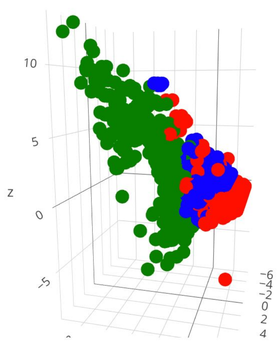
Previous work

- Our analysis builds on work which examines the characteristics of the data and finds trend including:
  - Donors generally donate to similar projects
  - Donors prefer projects near them
  - California has the highest rate of donors
  - Most only donate once or a small number of times

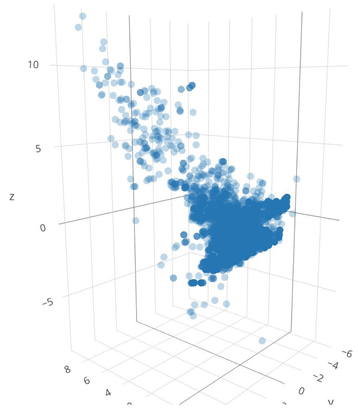
Methods

- Unsupervised learning to analyze data given
  - Principal Component Analysis used to reduce dimensions before clustering to group data in main sections
  - Latent Dirichlet Allocation on the textual data given to analyze descriptions and find similarities between essays
- Classification models on essays and textual data to understand factors contributing to project success
  - Bag of words and bigram representations of the project descriptions
    - Naive Bayes' Classifier
    - Random Forest Model
    - Support Vector Machines
  - Find words and bigrams that signal a project will be successful or result in out of state donations

Clustering with Principal Component Analysis



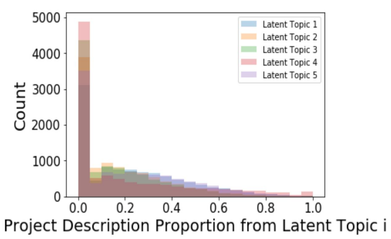
Three clusters applied to PCA graph

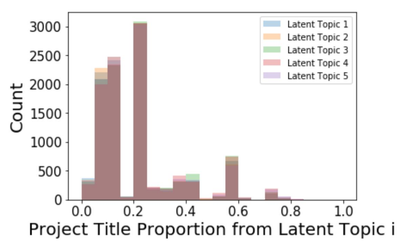


Data reduced in dimensionality using three principle components

Latent Dirichlet Allocation on Text Data

We ran LDA on text responses for "Project Title" and "Project Short Description"





Proportions of each latent topic contained in each document

Project Future Work

- More sophisticated analysis would be possible if we had access to the images displayed on the Donors Choose website by each project
- If we had time series data, we could model the probability of a project meeting its funding goal as a function of the current number of donors and funding level

Results of Classification Models

- Models performed better than random guessing, but were unable to achieve accuracies significantly above 60%
- Confusion matrix of whether a project is funded based on essay data (test set size = 2000, Random Forest Classifier, accuracy = 61.3%):

618	386
414	582

- Confusion matrix of whether a project receives out of state donations (test set size = 4000, Random Forest Classifier, accuracy = 56.0%):

1128	888
874	1110

Phrases Predicting Success

- Bigrams used to find phrases that indicated the success of projects in various classification models
- Top features that predict whether a project is funded or not:
  - 'abl hang'
  - 'learn unparallel'
  - '-- virtual'
  - 'ad stabil'
  - 'accept special'
- Top features that predict whether a project receives out of state donations:
  - 'applic skill'
  - 'better util'
  - 'read comput'
  - 'scienc elect'
  - 'accomplish dream'

Citations

- Shepard, James. DonorsChoose - Matching donors to causes.

Acknowledgments

We received guidance from Matthew Myers and Jonathan Lu at Office hours