CS451 Data Mining Project Proposal

1. **Project Title**

   Predict funding requests that deserve an A+(KDD Cup 2014)

2. **Team members**

   Yi Xiao 20293953

   Fengqing Li 20270195

3. **Project definition (KDD Cup 2014 Description)**

   DonorsChoose.org is an online charity that makes it easy to help students in need through school donations. At any time, thousands of teachers in K-12 schools propose projects requesting materials to enhance the education of their students. When a project reaches its funding goal, they ship the materials to the school.

   The 2014 KDD Cup asks participants to help DonorsChoose.org identify projects that are exceptionally exciting to the business, at the time of posting. While all projects on the site fulfill some kind of need, certain projects have a quality above and beyond what is typical. By identifying and recommending such projects early, they will improve funding outcomes, better the user experience, and help more students receive the materials they need to learn.

   For our project, we will approach the funding goal with supervised learning like logistic linear regression. In our data, we have both training and testing set for donations, essays, projects, resources and training data for outcomes. We can try to find the relationship between these data.

4. The approaches we are planning to use/implement

We are going to predict the A+ founding request by comparing data (donations, essays, projects, resources, outcomes).

Since the data we may use in: “donations.csv”, “essays.csv”, “projects.csv”, “outcomes.csv”, “sampleSubmission.csv” contains either quantitative and qualitative data, hence we may use both linear regression and logistic regression to predict funding requests that deserve an A+.

For example, we can make some linear regression between two of these data, and the then we can find some relationship by observing the plot of these data. And we will consider some variables like about the number of projects did the school have in the past, the number of great_chat projects for the teacher, the number of unique donors in this place, etc.

We will use some features for responses: great_chat, full_funded, one_non_teacher_referred_donor_giving_100_plus, etc.

5. Information about dataset

a. Data source

donations.csv - contains information about the donations to each project. This is only provided for projects in the training set.
essays.csv - contains project text posted by the teachers. This is provided for both the training and test set.
projects.csv - contains information about each project. This is provided for both the training and test set.
resources.csv - contains information about the resources requested for each project. This is provided for both the training and test set.
outcomes.csv - contains information about the outcomes of projects in the training set.
sampleSubmission.csv - contains the project ids of the test set and shows the submission format for the competition.

b. Data Size

The data is provided in a relational format and split by dates. Any project posted prior to 2014-01-01 is in the training set (along with its funding outcomes).

Donations.csv	Zip(252.91mb)
Outcomes.csv	Zip(11.18mb)
SampleSubmission.csv	Zip(793.65mb)
Resources.csv	Zip(193.83mb)
Essays.csv	Zip(402.27mb
Projects.csv	Zip(64.93mb)

Detail information about Data:

outcomes.csv
is_exciting - ground truth of whether a project is exciting from business perspective
at_least_1_teacher_referred_donor - teacher referred = donor donated because teacher shared a link or publicized their page
fully_funded - project was successfully completed
at_least_1_green_donation - a green donation is a donation made with credit card, PayPal, Amazon or check
great_chat - project has a comment thread with greater than average unique comments
three_or_more_non_teacher_referred_donors - non-teacher referred is a donor that landed on the site by means other than a teacher referral link/page
one_non_teacher_referred_donor_giving_100_plus - see above
donation_from_thoughtful_donor - a curated list of ~15 donors that are power donors and picky choosers (we trust them selecting great projects)
great_messages_proportion - how great_chat is calculated. proportion of comments on the project page that are unique. If > avg (currently 62%) then great_chat = True
teacher_referred_count - number of donors that were teacher referred (see above)
non_teacher_referred_count - number of donors that were non-teacher referred (see above)

projects.csv
projectid - project's unique identifier
teacher_acctid - teacher's unique identifier (teacher that created a project)
schoolid - school's unique identifier (school where teacher works)
school_ncesid - public National Center for Ed Statistics id
school_charter - whether a public charter school or not (no private schools in

the dataset)
school_magnet - whether a public magnet school or not
school_year_round - whether a public year round school or not
school_nlns - whether a public nlns school or not
school_kipp - whether a public kipp school or not
school_charter_ready_promise - whether a public ready promise school or not
teacher_prefix - teacher's gender
teacher_teach_for_america - Teach for America or not
teacher_ny_teaching_fellow - New York teaching fellow or not
primary_focus_subject - main subject for which project materials are intended
primary_focus_area - main subject area for which project materials are intended
secondary_focus_subject - secondary subject
secondary_focus_area - secondary subject area
resource_type - main type of resources requested by a project
poverty_level - school's poverty level.
highest: 65%+ free of reduced lunch
high: 40-64%
moderate: 10-39%
low: 0-9%
grade_level - grade level for which project materials are intended
fulfillment_labor_materials - cost of fulfillment
total_price_excluding_optional_support - project cost excluding optional tip that donors give to DonorsChoose.org while funding a project
total_price_including_optional_support - see above
students_reached - number of students impacted by a project (if funded)
eligible_double_your_impact_match - project was eligible for a 50% off offer by a corporate partner (logo appears on a project, like Starbucks or Disney)
eligible_almost_home_match - project was eligible for a $100 boost offer by a corporate partner
date_posted - data a project went live on the site

donations.csv
donationid - unique donation identifier
projectid - unique project identifier (project that received the donation)
donor_acctid - unique donor identifier (donor that made a donation)
donor_city
donor_state
donor_zip
is_teacher_acct - donor is also a teacher
donation_timestamp
donation_to_project - amount to project, excluding optional support (tip)
donation_optional_support - amount of optional support
donation_total - donated amount

dollar_amount - donated amount in US dollars
donation_included_optional_support - whether optional support (tip) was included for DonorsChoose.org
payment_method - what card/payment option was used
payment_included_acct_credit - whether a portion of a donation used account credits redemption
payment_included_campaign_gift_card - whether a portion of a donation included corporate sponsored giftcard
payment_included_web_purchased_gift_card - whether a portion of a donation included citizen purchased giftcard (ex: friend buy a giftcard for you)
payment_was_promo_matched - whether a donation was matched 1-1 with corporate funds
via_giving_page - donation given via a giving / campaign page (example: Mustaches for Kids)
for_honoree - donation made for an honoree
donation_message - donation comment/message. Used to calcualte great_chat

essays.csv
projectid - unique project identifier
teacher_acctid - teacher id that created a project
title - title of the project
short_description - description of a project
need_statement - need statement of a project
essay - complete project essay

resources.csv
resourceid - unique resource id
projectid - project id that requested resources for a classroom
vendorid - vendor id that supplies resources to a project
vendor_name
project_resource_type - type of resource
item_name - resource name (ex: ipad 32 GB)
item_number - resource item identifier
item_unit_price - unit price of the resource
item_quantity - number of a specific item requested by a teacher

6. All software we are going to use

Python(Maybe), R, RStudio, Github, Microsoft Office

7. Timeline for Project

Item	Description	todo by Date	Status
1	Meet with team, decide on topic, features and target variable(s), split tasks	09/26	Completed
2	Download data, setup git repo with team, install R packages	09/30	Completed
3	Data exploration / preprocessing	10/03	Completed
	Project Proposal Due	**10/03**	Completed
4	Data preprocessing	10/10	TBC
5	Feature selection / creation	10/17	TBC
6	Run models, evaluate accuracy, plot accuracy results (Round 1)	10/24	TBC
7	Update models, run cross validation, evaluate accuracy, plot accuracy results (Round 2)	10/31	TBC
8	Round 3 (with significance of results always tested)	10/03	TBC
	Project Progress Report Due	**11/03**	TBC
9	Plot comparisons: models (m1, m2, m3, ...), baselines (random), compute % improvements	11/07	TBC
...	(Challenges in your project, feature transformations, model majority voting, MapReduce, ...)	...	...
	Project Presentation	**12/05**	TBC
	Project Report and Code Due	**12/07**	TBC