Project Progress Report

1) **Project title**

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

2) **Team members**

Yi Xiao 20293953
Fengqing Li 20270195

3) **Timeline**

Week 7-8

Download data frome KDD Cup 2014 website

Setup github repo with team

Install R packages for preparation

Week 9

Data preprocessing

Feature selection/creation

Choose data from official dataset of KDD Cup 2014 to make our own dataset

Determine target variable and features for our own data

Week 10

Find low p-value by running logistic regression, throw away all other columns in our dataset.

Repeats previous to other official datasets to get our own data.

Week 11

Run different models, try to include interactions between variables, cross validate your results to get more accurate test error, plot accuracy results.

(Round 1 for training1.csv)

Writing Project Progress Report

4) **Problem definition**:

The 2014 KDD Cup asks participants to identify projects that are exceptionally exciting to the business.

Target variable: is_exciting - ground truth of whether a project is exciting from business perspective

"Exciting" Projects

To be exciting, a project must meet all of the following five criteria. The name in parentheses indicates the field containing each feature in the data set.

was fully funded

had at least one teacher-acquired donor

has a higher than average percentage of donors leaving an original message

has at least one "green" donation

has one or more of :

- donations from three or more non teacher-acquired donors

- one non teacher-acquired donor gave more than $100

- the project received a donation from a "thoughtful donor"

5)**Data Analysis:**

i) Features/predictors that we are using:

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)

school_latitude, school_longitude, school_zip, school_metro,

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

poverty_level - school's poverty level.

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

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

donor_zip, is_teacher_acct - donor is also a teacher

donation_to_project - amount to project, excluding optional support (tip)

donation_optional_support - amount of optional support

donation_total - donated amount

donation_included_optional_support - whether optional support (tip) was included for DonorsChoose.org

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

ii) Why you think they are informative/why you think that these features affect the target variable(is_exciting):

Quantitative features (p-values, correlation results):

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0642  -0.6061  -0.4733  -0.2063   3.7675

Coefficients: (1 not defined because of singularities)
                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                        -1.208e+00  4.276e-01  -2.824  0.00474 **
donor_zip                           4.583e-07  1.140e-06   0.402  0.68769
donation_to_project                -2.811e-03  1.049e-03  -2.680  0.00737 **
donation_optional_support           9.910e-04  6.137e-03   0.161  0.87172
donation_total                             NA         NA      NA       NA
school_latitude                    -5.190e-03  6.915e-03  -0.751  0.45292
school_longitude                    2.122e-04  5.798e-03   0.037  0.97081
school_zip                         -2.379e-06  3.294e-06  -0.722  0.47022
total_price_excluding_optional_support  1.244e-01  7.627e-03  16.309  < 2e-16 ***
total_price_including_optional_support -1.057e-01  6.484e-03 -16.303  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7378.3  on 9999  degrees of freedom
Residual deviance: 6749.9  on 9991  degrees of freedom
AIC: 6767.9

Number of Fisher Scoring iterations: 7
```

We choose donation_to_project, total_price_excluding_optional_support and total_price_including_optional_support as our features to continue analysis with our target value – is_exciting.

6) **Models / Prediction Algorithms**: List the models you are considering and explain why you believe that these are a good selection that fits the nature of your problem. My target variable is a class with K > 2 possible values hence it makes sense to use LDA.

```
Call:
lda(is_exciting ~ donation_to_project + total_price_excluding_optional_support +
    total_price_including_optional_support, data = train1)

Prior probabilities of groups:
    f     t
0.879 0.121

Group means:
  donation_to_project total_price_excluding_optional_support
f           114.72116                                557.0586
t            58.62131                                579.2747
  total_price_including_optional_support
f                               663.0769
t                               681.6503

Coefficients of linear discriminants:
                                                LD1
donation_to_project                     -0.0006817934
total_price_excluding_optional_support   0.0594443599
total_price_including_optional_support -0.0501862855
```

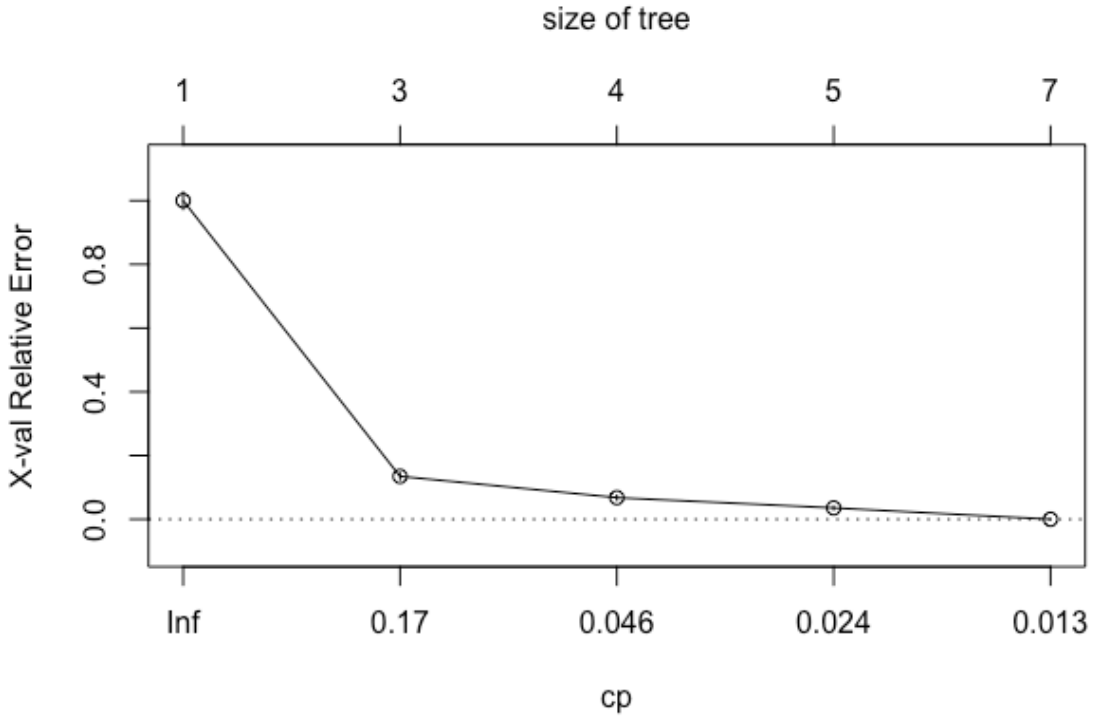
Binary features(tree):

```
Variables actually used in tree construction:
[1] at_least_1_green_donation                     at_least_1_teacher_referred_donor
[3] fully_funded                                  great_chat
[5] one_non_teacher_referred_donor_giving_100_plus three_or_more_non_teacher_referred_donors

Root node error: 1210/10000 = 0.121

n= 10000

        CP nsplit rel error   xerror      xstd
1 0.432645      0  1.000000 1.000000 0.0269527
2 0.066942      2  0.134711 0.134711 0.0104650
3 0.032231      3  0.067769 0.067769 0.0074530
4 0.017769      4  0.035537 0.035537 0.0054077
5 0.010000      6  0.000000 0.000000 0.0000000
```



We will continue to model our data in trees and make plots to show more details about the dataset in the next few weeks.

7) **Rounds / Iterations** you are considering to improve accuracy. At least one round of modeling should be run and included in this report. If your team has tried more approaches, include them here.

We have three data files to improve the accuracy of our project: training1.csv, training2.csv and traininig3.csv. We used the training1.scv file to build our model firstly. We also tried different sets of features to train our model (features have different p-values) and features that have low p-value make our model better.

We will run all datasets in the rest of the semester and we will try model like trees to compare different models and to improve accuracy.

Since the KDD Cup 2014 did not provide the test set of all their data, we will work with training data and split them into train-test parts. Treat test data as unknown.

8) Describe any challenges you have met so far.

The most challenge in our project is the super large dataset form KDD Cup 2014, since the size of all data is almost 3GB. We cannot normally run these data on our computers, our RStudio becomes extremely slow when loading these datasets.

In total, we have three datasets and there are more than 600,000 data in two datasets and there are more than 1000,000 data in another dataset. To deal with these huge dataset, we choose 30,000 data from the official dataset provided by KDD Cup 2014 and we combine whole variables of those datasets to make our own data.

And other difficulties in our project like coding in RStudio for our model, finding libraries and a large amount of variables we need to calculate, etc. We have already solved these problems by self-learning and searching information we think.