Final Project Report

1) **Project title**

Predicting Exciting Project at DonorsChoose.org (KDD Cup 2014)

2) **Team members**

Yi Xiao 20293953
Fengqing Li 20270195

3) **Timeline**

Week 12

Modify model features (select features by different range, deal with outlier and high test error)

Week 13

Data preprocessing for LR, LDA, Tree, Random Forest (5-7 round)

Week 14

Calculate MSE, Mean Test error of all models

Week 15

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

Combine all codes, models, results.

Round 7-10 for all models

Writing Project Presentation

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)

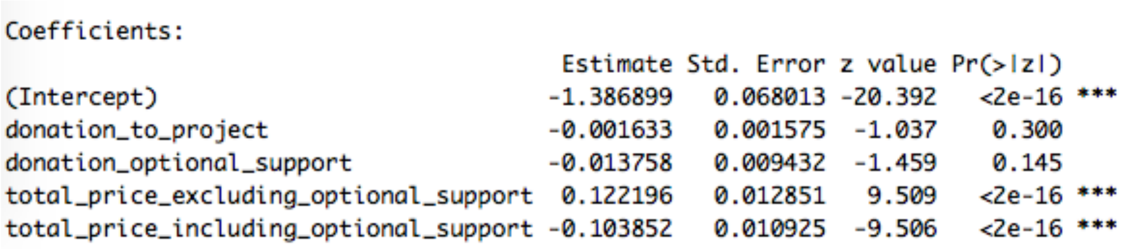
payment_was_promo_matched - whether a donation was matched 1-1 with corporate funds

via_giving_page - donation given via a giving / campaign page (example: Mustaches for Kids)

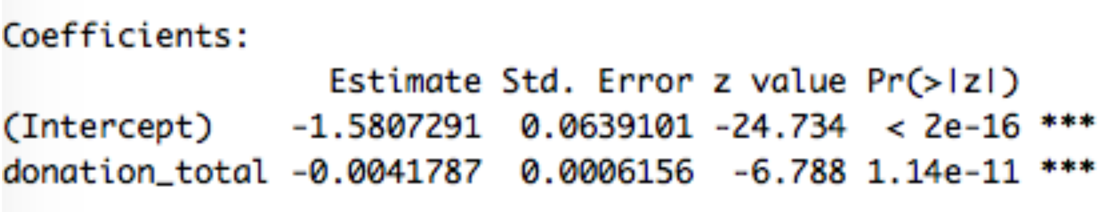
for_honoree - donation made for an honoree

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

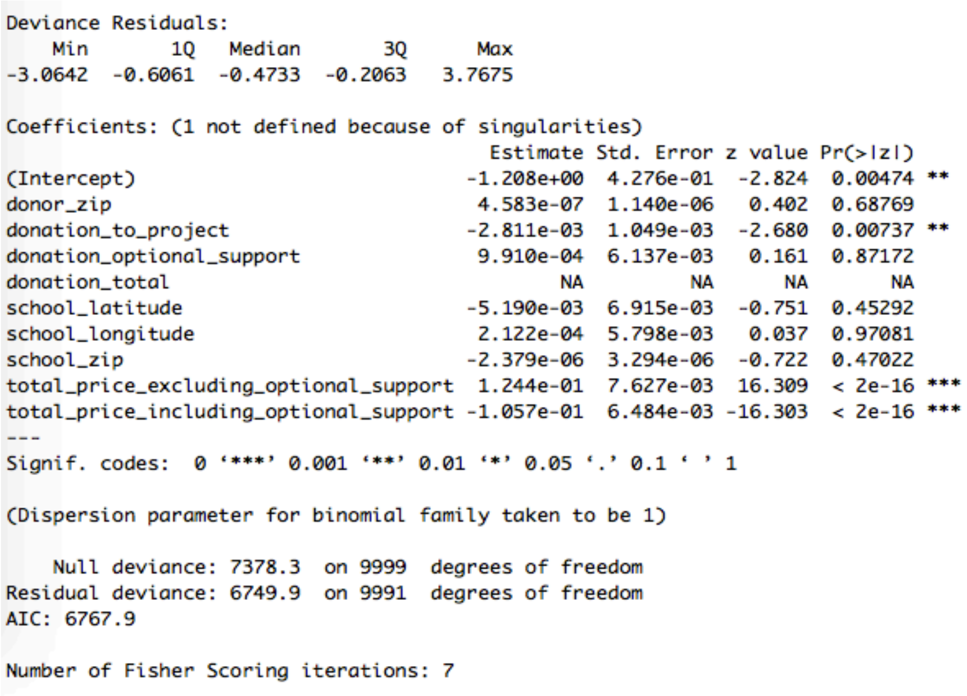
Quantitative features (p-values, correlation results):



```
Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)                         -1.386899   0.068013 -20.392   <2e-16 ***
donation_to_project                 -0.001633   0.001575  -1.037    0.300    
donation_optional_support           -0.013758   0.009432  -1.459    0.145    
total_price_excluding_optional_support  0.122196   0.012851   9.509   <2e-16 ***
total_price_including_optional_support -0.103852   0.010925  -9.506   <2e-16 ***
```



```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)    
(Intercept)   -1.5807291  0.0639101 -24.734  < 2e-16 ***
donation_total -0.0041787  0.0006156  -6.788 1.14e-11 ***
```



```
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_total, total_price_excluding_optional_support and total_price_including_optional_support as our features to continue analysis with our target value – is_exciting.

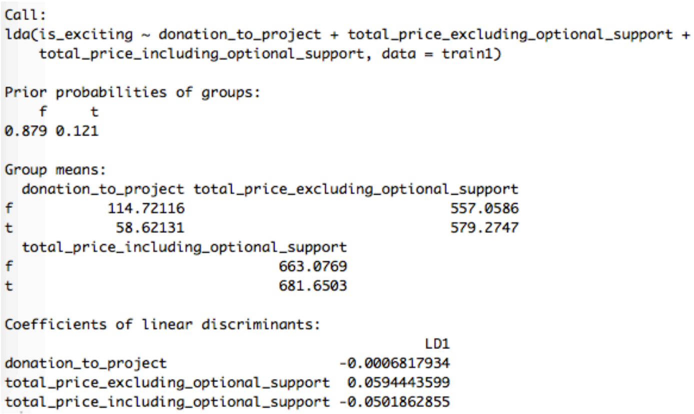
6) **Models / Prediction Algorithms**:

YiXiao’s project models :

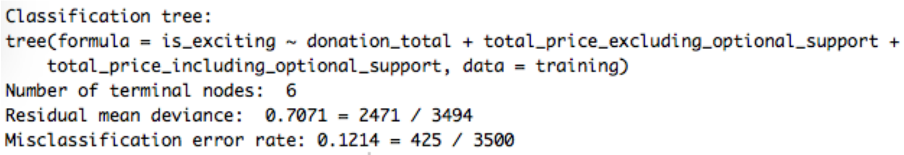
1. Only for quantitative features (5 features)
   a. Donation_to_project - amount to project, excluding optional support
   b. Donation_optional_support - amount of optional support
   c. Donation_total - donation amount
   d. Total_price_excluding_optional_support - project cost excluding optional tip
   e. Total_price_including_optional_support - project cost excluding optional tip
2. Donation, poverty level and total price (4 features)
   a. donation_to_project
   b. donation_optional_support
   c. donation_total
   d. Poverty_level - school's poverty level (highest, high, moderate, low)

3. Everything except donation and total price (26 features)
   a. Is_teacher_acct - donor is also a teacher

b. For_honoree - donation made for an honoree
c. at_least_1_teacher_referred_donor - teacher referred = donor donated because teacher shared a link or publicized their page
d. fully_funded - project was successfully completed
e. teacher_referred_count - number of donors that were teacher referred
f. great_chat - project has a comment thread with greater than average unique comments
g. …

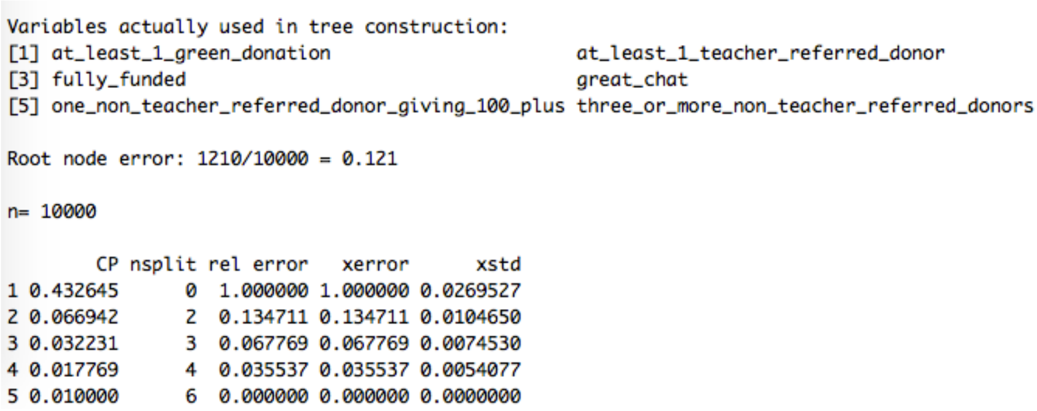


```
Call:
lda(is_exciting ~ donation_to_project + total_price_excluding_optional_support +
    total_price_including_optional_support, data = train1)

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
```

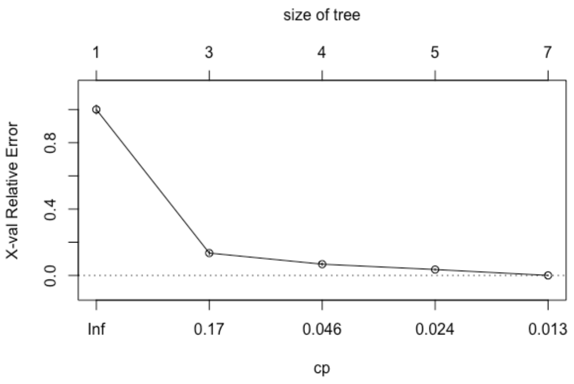


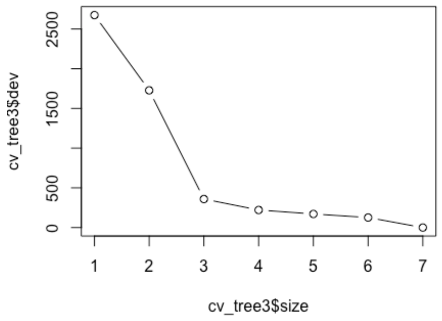
```
Classification tree:
tree(formula = is_exciting ~ donation_total + total_price_excluding_optional_support +
    total_price_including_optional_support, data = training)
Number of terminal nodes:  6 
Residual mean deviance:  0.7071 = 2471 / 3494 
Misclassification error rate: 0.1214 = 425 / 3500 
```

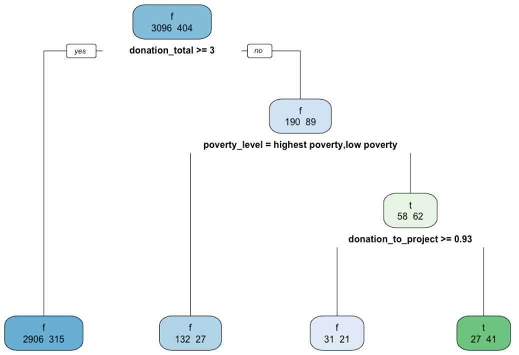
Binary features(tree):

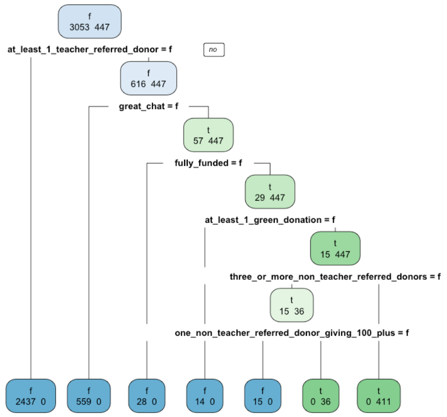


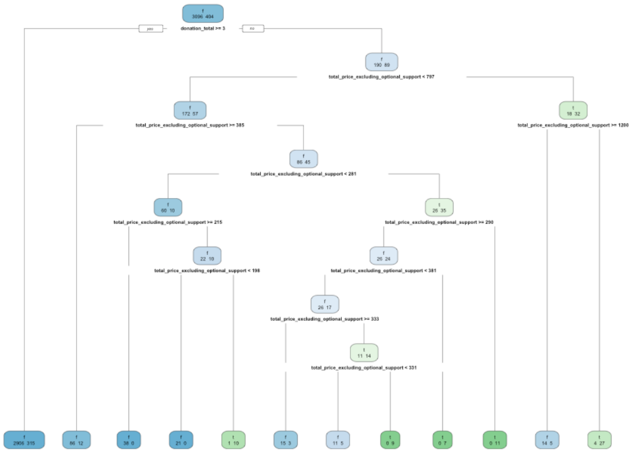
```
Variables actually used in tree construction:
[1] at_least_1_green_donation                  at_least_1_teacher_referred_donor
[3] fully_funded                               great_chat
[5] one_non_teacher_referred_donor_giving_100_plus three_or_more_non_teacher_referred_donors

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
```








f
3053 447
at_least_1_teacher_referred_donor = f
no
f
616 447
great_chat = f
t
57 447
fully_funded = f
t
29 447
at_least_1_green_donation = f
t
15 447
three_or_more_non_teacher_referred_donors = f
t
15 36
one_non_teacher_referred_donor_giving_100_plus = f
f
2437 0
f
559 0
f
28 0
f
14 0
f
15 0
t
0 36
t
0 411


f
3096 404
yes
donation_total >= 3
no
f
190 89
total_price_excluding_optional_support < 797
f
172 57
total_price_excluding_optional_support >= 385
t
18 32
total_price_excluding_optional_support >= 1200
f
86 45
total_price_excluding_optional_support < 281
f
60 10
total_price_excluding_optional_support >= 215
t
26 35
total_price_excluding_optional_support >= 290
f
22 10
total_price_excluding_optional_support < 198
f
26 24
total_price_excluding_optional_support < 381
f
26 17
total_price_excluding_optional_support >= 333
t
11 14
total_price_excluding_optional_support < 331
f
2906 315
f
86 12
f
38 0
f
21 0
t
1 10
f
15 3
f
11 5
t
0 9
t
0 7
t
0 11
f
14 5
t
4 27

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.

I have 10 data files to improve the accuracy of our project: 1.csv, 2.csv and 3.csv... 10.csv file. I used every file to build my 3 main models. Ie also tried different sets of features to train my model (features have different p-values) and features that have low p-value make my model better.

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.

Training data: 70% of 5000 data.

Testing data: 30% of 5000 data.

8) Prediction accuracy
Accuracy:
Yi Xiao(LR, LDA, Tree, Random Forest):
Model 1: 43.5%
Model 2: 62.2%
Mode 3: 67.4%
Mean Test Error:
Model 1: 14.5%
Model 2: 10.7%
Mode 3: 7.4%

9) 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 larger than 3GB. I cannot normally run these data on my computers, my RStudio becomes extremely slow when loading those datasets.

In total, I have ten datasets and there are more than 600,000 data in every dataset and there are more than 1000,000 data in another dataset. To deal with these huge dataset, the data preprocessing is so important for me.

Also, high mean test error is also a big challenge since I did not have the test data from KDD website. My own test datasets may not accurate, this can make my test error high.

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

This is my first data class and I learned a lot in this semester and the challenge for this class also is a great experience for me. I will learn more about data in the future and I really enjoy this semester.