Name: Eagle Xuhui Ying WFU ID: 06648543

1. Introduction

DonorsChoose.org (https://www.donorschoose.org/) is an online charity helping students who need it 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. While all projects on the site fulfill some need, specific projects have a quality above and beyond typical, which are designated "exciting projects". By identifying and recommending such projects early, Donors Choose would like to be able to predict which projects are most likely to be exciting projects before the donation period has closed. Also, Donors Choose would like to understand its donors better to help them improve the overall user experience for donors who visit its website and encourage repeat donations. There are two given datasets. The DonorMerge_Final.csv (328018 rows, 40 columns) contains information on thousands of projects posted on DonorsChoose.org. These indicators of whether or not it was identified as an "exciting project". Another file, Donations.csv (1048575 rows, 21 columns), contains information on individual donors to DonorsChoose.org.

2. Data Quality Issues

DonorMerge_Final.csv (donormerge):

(1) Use 'lubridate' package to transform date_posted to days_from_now (the time difference between the posted date to today)
(2) Remove unique 'id' columns: projectid, teacher_acctid, schoolid, school_ncesid
(3) Remove categorical variables with too many categories: projectid, teacher_acctid, schoolid, school_ncesid, school_city, school_state, school_zip, school_county, school_metro, school_district, secondary_focus_subject, primary_focus_subject
(4) Drop columns with more than 20% missing values: secondary_focus_subject, secondary_focus_area
(5) Transform all categorical variables into factors:
one_non_teacher_referred_donor_g, school_charter, school_magnet, school_year_round, school_nlns, school_kipp, school_charter_ready_promise, teacher_teach_for_america, teacher_ny_teaching_fellow, eligible_double_your_impact_matc, eligible_almost_home_match
(6) Make a recipe, specify a formula, impute mean/mode to numerical/categorical missing values, normalize the numeric variables, and dummy encode nominal predictors

Donations.csv (donations):

(1) Remove unique 'id' columns: donationid, projectid, donor_acctid
(2) Impute median to missing values of three numerical columns: donation_to_project, donation_optional_support, donation_total
(3) Remove categorical variables with too many categories: donor_city, donor_state, donor_zip, donation_timestamp, donation_message
(4) Use na.omit() to remove the whole rows of other missing values

(5) Use 'filter' to remove outliers: donation_to_project >= 500, donation_optional_support >= 100, donation_total >= 300
(6) Create dummy variables: is_teacher_acct, dollar_amount, donation_included_optional_support, payment_method, payment_included_acct_credit, payment_included_campaign_gift_card, payment_included_web_purchased_gift_card, payment_was_promo_matched, via_giving_page, for_honoree

3. Identifying Exciting Projects

Business focused presentation of the final analytical solution

After using logistic regression, neural network, and random forest, the random forest model is the best to predict exciting projects. After I conducted variable importance, I found ***teacher_referred_count, great_messages_proportion, non_teacher_referred_count, one_non_teacher_referred_donor_g, days_from_now*** are the five most essential variables in my random forest model. Then, we can put these variables back into the logistic regression model to see their influence on the log odds of the logistic regression model.

Discussion of key factors that can be used to identify exciting projects

(1) teacher_referred_count: When the number of teacher-referred donors increases, the likelihood of a project being exciting will also increase. Also, as shown by the model, a teacher's referral has a better effect than a non-teacher referral.
(2) great_messages_proportion: The higher proportion of unique comments on a page, the more likely a donation project will be regarded as exciting.
(3) non_teacher_referred_count: Increasing the number of referrals not from teachers can also increase the excitement of a project.
(4) one_non_teacher_referred_donor_g: If a donor landed on the site by means other than a teacher referral/link, the probability of project excitement would increase.
(5) days_from_now: The earlier a project goes live on the site, the less likely it will be regarded as exciting.

4. Understanding Donors

Business focused overview of the final analytical solution

I used K-Means Clustering to divide all 990949 donors into 5 clusters, with 394947, 104110, 52427, 293621, and 145844 donors in each group. In the following part of this report, I call them Group 1, Group 2, Group 3, Group 4, and Group 5 separately.

Discussion of key characteristics of the different types of donors

(1) Group 1: It has the largest number of donors. Their donation amounts are all less than $100. However, no cash is received. The proportion of donating corporate-sponsored gift cards is the highest among all five groups.
(2) Group 2: Their donation amounts all exceed $100.
(3) Group 3: It has the smallest number of donors. However, they donated the largest amount of money to schools, with the largest amount of optional tips for each person. No donors are teachers in this group. Their donation amounts are all greater than $100. All donators got optional support. No donations included a gift card purchased by the

donor.
(4) Group 4: Their donations are less than $100. Compared to other groups, this group has a minor portion of donations used accounts credit redemption. The proportion of donating corporate-sponsored gift cards is the lowest among all five groups. No donations included a gift card purchased by the donor. Most donations were given through a giving/campaign page.
(5) Group 5: Their donation amounts all lie between $10 and $100.

5. Recommendation

(1) Since receiving referrals either from teachers or not from teachers can both increase the project excitement, Donors Choose should recommend projects with many referrals early to improve funding outcomes. Also, they should pay more attention to the number of teachers' referrals because they are more influential than non-teacher referrals when predicting project excitement.
(2) Donors Choose should focus more on the project that recently went live on the site because they are more likely to be exciting projects than those posted earlier. Also, they should pay close attention to unique comments on a page, which is an essential indicator of an exciting donation project.
(3) A large proportion of donors are not desired to donate, they donate money less than $100 each time, and no cash is received in this process. While it is nearly a waste of time to spend time and energy on these donors, we should focus on donors who donate more money at once. As to donors who donate more than $100 each, we need to encourage repeat donations by keeping in touch with them and using giving/campaign pages. Likely, their richness and willingness can help Donors Choose to increase the donations they receive.
(4) Methods of donations are also an important factor to consider while conducting donor segmentation. Donors who donate with cash and tips tend to donate more amount of money at one time. By comparison, donations using corporate-sponsored gift cards or account credit redemption are usually associated with small amounts and unwillingness to donate. Therefore, Donors Choose needs to keep in contact with donors who like to contribute with cash and try their best to encourage them to donate again.
(5) As for Donors Choose, they should do more thorough investigations to make sure the completeness of surveys. For example, in my final model, the proportion of unique comments on a page, the number of teacher-referred donors, and the number of non-teacher-referred donors are all significant predictors of the model. However, there are a lot of missing values in these variables. Imputing the median to these missing values will probably decrease the accuracy and credibility of my classification result. A more complete and informative dataset can improve the results of these types of analyses in the future.