1. Introduction to the business problem

An internet organization called DonorsChoose.org makes it simple to support underprivileged students by collecting school donations. Many projects were proposed by numerous K–12 teachers constantly to ask for materials to improve their kids' education.

Besides fulfilling the foundational requirements of children, some projects provide more benefits to them, regarded as ‘exciting’ projects. To improve the funding effects, DonorsChoose.org expects to identify those projects early. In addition, DonorsChoose.org would like to understand its donors better to enhance its customer service and stimulate more donations.

2. Data discussion

2.1.Data overview

The data for projects is named as ‘donation’ and the data for donors is named as ‘donor’. There are over 10M rows in ‘donor’ and there are over 0.3M rows in ‘donation’ .

2.2.Data quality

The missing values problem is more serious in ‘donation’ and these problems have been dealt with mean and mode imputation.

Besides, exploratory analysis assists in identifying outliers in some numeric variables. The box plots (see Appendix 1) show that the vast majority of data points are concentrated in small values, severely out of normal distribution. Therefore, after trying different cutoff values to make the distribution normalize, some data rows were deleted which are higher than the cutoff values. Simultaneously, the dataset remained over 95% rows.

3. Presentation of the final solution addressing the first goal of identifying exciting projects

3.1.Business-focused presentation of the final analytical solution

Eventually, the logistic regression model, the logistic regression reduced model (removing variables manually), and a random forest model were used to predict the exciting programs. By comparing the major evaluation metrics used in this project, accuracy, and ROC_AUC, the Random Forest model with the highest accuracy and ROC_AUC (see Appendix 2) is decided as the final model to predict exciting programs.

The final predictive model, the Random Forest Model has an accuracy of 99.47%, a ROC_AUC of 99.90%, a recall of 99.47%, and a precision of 99.96%. 99.47% accuracy means that 99.47% of data has been classified correctly; 99.9% ROC_AUC means that the predictive model has a very strong ability to classify FALSE and TRUE; 99.47% recall means that among the data predicted to be positive, 99.47% of the data is really positive; 99.96% precision means that 99.96% positive data has been found (predicted as positive).

3.2.Discussion of key factors

Some variables have a higher importance in influencing the predictive model (see Appendix 4). Generally speaking, the projects, which are released more recently and are eligible for discounts provided by a corporate partner, with more messages on the website for schools, with a higher cost of fulfillment, for schools located in higher latitudes, with lower costs of donation and higher optional tips, are more exciting. In addition, the programs whether existing are also depending on the schools’ states.

4. Presentation of the final solution addressing the second goal of understanding donors

4.1. Business-focused overview of the final analytical solution

The k-means clustering algorithm is used in this clustering analysis. And the elbow method is used to determine the optimal k value, which is 5 in this project. Therefore, the donors were divided into 5 groups. 5 clusters of sizes are 69317, 24723, 180648, 196166, and 86423, which indicates the good performance of clustering.

4.2. Discussion of key characteristics of donors

By checking the distribution of each variable in different clusters, some variables allow different clusters to have distinct features were found. 5 clusters have the following own characteristics:

- Cluster 1:

The percentage of teacher donors is higher than average; the dollar amounts for the project excluding optional tips, optional tips, and the total amount are all higher than average; the major payment methods are credit cards.

- Cluster 2:

The percentage of teacher donors is higher than average; all dollar amounts are the highest among the clusters; the major payment methods are campaigns of cooperation partners and credit cards.

- Cluster 3:

The proportion of teacher donors is less than average; all the dollar amounts are average; the majority of donors did not really pay; the percentage of users donating with gift cards is the highest; the proportion of donors entering the website through the campaign page is the lowest.

- Cluster 4:

The proportion of teacher donors is the least; all dollar amounts are average; the payment methods are similar to the general donors; the percentage of users donating with gift cards is the lowest; the proportion of donors entering the website through the campaign page is the highest.

- Cluster 5:

The proportion of teacher donors is the highest; all dollar amounts are the lowest; the major payment methods are non-cash received, PayPal, and credit cards; the percentage of users donating with gift cards is higher than average; the proportion of donors entering the website through the campaign page is the highest.

5. Summary & Conclusion

5.1. Recommendations on predicting exciting programs and encouraging donation

- **Predicting**

Some factors with a higher importance in predicting and with higher interpretability under the business context should be regarded as the most important features. It is suggested that the donation projects with more unique and serious comments on pages, higher cost of fulfilling those projects, releasing more recently, more support offered by cooperation partners, and more tips donated by donors are more exciting projects.

- **Encouraging donation**
  - **Joint Charity Marketing**

According to the predictive model, it is found that the projects with more support from cooperation partners are more likely to be exciting programs with more donations. Therefore, some joint charity marketing campaigns can be held to encourage more people to donate more. For instance, DonorsChoose.org can jointly launch some products with Disney (similar companies may be related to children because they can inspire more sympathy), and the general price of the products will be donated to these projects. At the same time, those who buy these charitable products can receive an additional reward such as a charity medal or coupons for Disney tickets.

- **Low cost and tipping encouraged**

Because the projects involved in low-cost items were found to be more exciting, it is suggested that DonorsChoose.org proposes more low-cost projects that are easy to be donated by users. In addition to basic costs, DonorsChoose.org could use animations or stories about donation projects to encourage donors to give more tips when giving, by inspiring compassion.

5.2.Recommendations on donor segments

In terms of different segments, different strategies should be used to improve user experience and encourage donations:

- Cluster 1:

Since credit cards are the primary means of payment for this group, work with credit card companies to promote donations by offering proportional cashback.

- Cluster 2: the most valuable segments

DonorsChoose.org could establish a reward mechanism to thank them for their donations and encourage their repeat donations. For instance, it can present a charity trophy to donors after they have donated five times.

- Cluster 3: gift card groups

Donors in this segment most like to donate with gift cards, so DonorsChoose.org can give another gift card so that donors can buy something and donate again.

- Cluster 4: users from corporate partners

Since most members of this group come from the activity page of the partner, this organization can launch the activity discounts together with the partner to encourage real donations.

- Cluster 5: ignorable

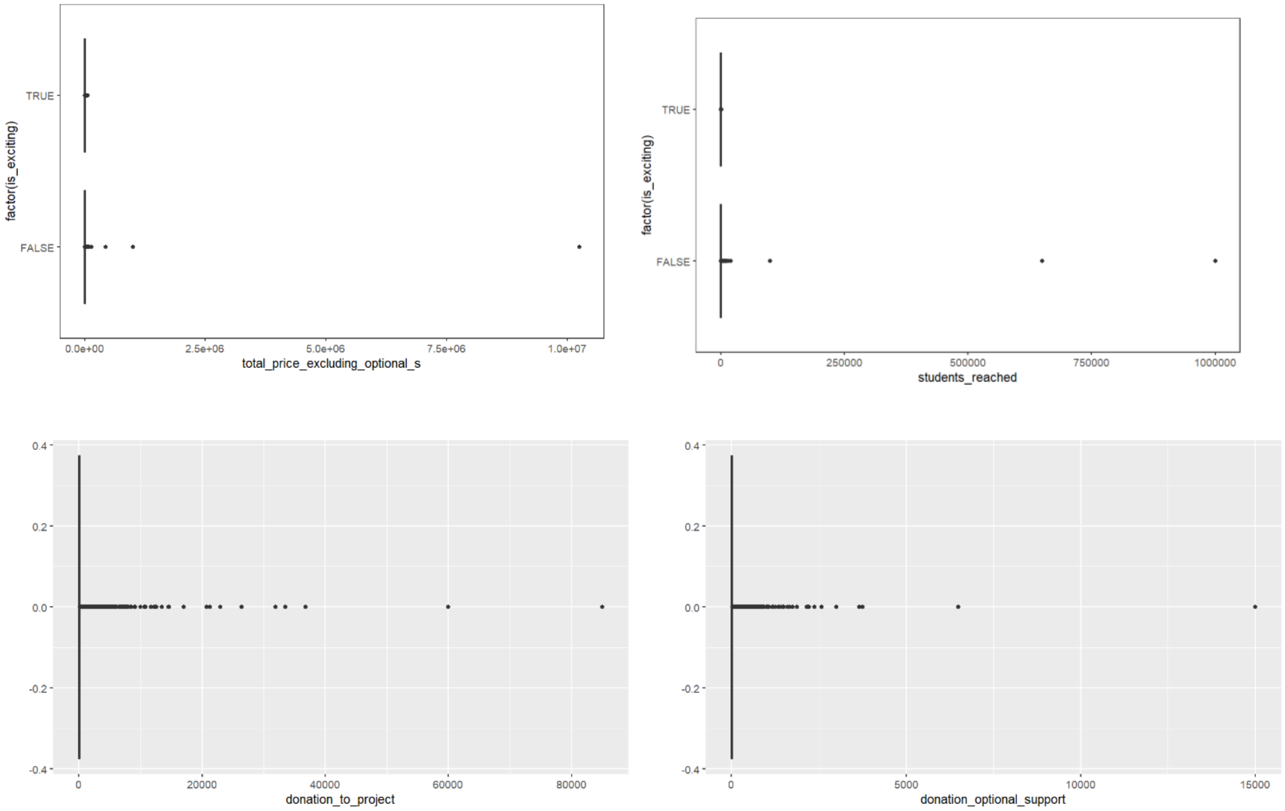
5.3.Recommendations on improving analysis results

Because this data set has serious data problems, especially missing values, DonorsChoose.org can improve data quality (accuracy and completeness) later in the process of collecting data. And it can add some variables appropriately to better help predict good projects and understand donors.

6. Appendix

Appendix 1

Box plots for numeric variables with outliers



Appendix 2

Table for evaluating three predictive models

	part	.estimator	accuracy	roc_auc
LR	testing	binary	91.22%	93.88%
	train	binary	91.19%	93.83%
LR reduced	testing	binary	91.23%	93.89%
	train	binary	91.18%	93.80%
RF	testing	binary	99.50%	99.88%
	train	binary	99.47%	99.90%

Appendix 3

Table of precision and recall of final RF model

.metric	.estimator	.estimate	part
recall	binary	0.994437	training

recall	binary	0.994713	testing
precision	binary	0.999566	training
precision	binary	0.999603	testing

Appendix 4

Important Variables

Variable	Importance	Sign
great_messages_proportion	120.4121	POS
teacher_referred_count	114.799	POS
non_teacher_referred_count	67.59212	POS
one_non_teacher_referred_donor_g_TRUE.	50.99597	POS
fulfillment_labor_materials	43.86589	POS
days	40.82673	NEG
eligible_double_your_impact_matc_TRUE.	13.24791	POS
teacher_teach_for_america_TRUE.	12.90355	POS
eligible_almost_home_match_TRUE.	12.54855	POS
school_latitude	11.29862	POS
school_state	9.451699	POS
total_price_excluding_optional_s	5.561931	NEG
total_price_including_optional_s	5.272428	POS