Predicting Classroom Project Funding

1 Introduction

Across the country, many public school teachers use their own money to buy supplies for their classrooms, including basic essentials like notebooks and pencils. If schools and teachers cannot afford these supplies, students may miss out on learning opportunities. DonorsChoose.org was established to help teachers get the resources they need for classroom projects. Through the DonorsChoose.org website, teachers can request materials for their classrooms. Donors can browse through requests, and select a project they would like to support financially.

In 2015, 72% of classroom projects posted on DonorsChoose.org were successfully funded. By analyzing data available from DonorsChoose.org, the goal of this capstone project is to understand what types of projects receive full funding and use this information to help teachers improve their chances of success.

2 Datasets

Data was obtained from the DonorsChoose.org open datasets:

https://research.donorschoose.org/c/opendata

Two datasets were selected for analysis:

- Project data (opendata_projects000.gz), which includes information on every project posted to the website, including school location, type of school, teacher attributes, project categories, project pricing and funding status.
- Essay data (opendata_essays000.gz), which includes the full text of the teacher-written requests that appear on the website, plus the thank you note and impact letter that are written after a project is completed.

3 Overall Approach

The two datasets were imported into R (version 3.3.1), cleaned and then joined based on project ID. Exploratory data analysis was conducted, including geographic and sentiment analysis. The scope was limited to data from 2015, with the intention of selecting recent data that spans an entire year, but does not include any live projects (as is the case for 2016).

The data was split into training and test sets. A classification tree was constructed using the training set, and assessed using the test set. The predictive model was used to make recommendations for teachers using DonorsChoose.org.

This report outlines the data import, data cleaning, exploratory data analysis, and classification tree building stages. It concludes with recommendations for the teacher, and suggestions for future work with the DonorsChoose.org datasets.

4 Data Import

Each dataset .gz file was unzipped to extract a .csv file. In R, the fread command was used to import the data, as it is known to be faster than read.csv for large datasets. Since the essay data file is larger than 3 GB, the import was limited to the columns of interest (including project ID and essay), and the unneeded columns were dropped during import (including thank you letter and impact letter). Since the csv files did not include a header, the variable names were specified in the import command, as shown in the following code.

```
project <- fread("opendata_projects000/opendata_projects000.csv", sep = ",",
  header = FALSE, data.table = FALSE, col.names = c('_projectid', '_teacher_acctid',
  '_schoolid', 'school_ncesid', 'school_latitude', 'school_longitude', 'school_city',
  'school_state', 'school_zip', 'school_metro', 'school_district', 'school_county',
  'school_charter', 'school_magnet', 'school_year_round', 'school_nlns', 'school_kipp',
  'school_charter_ready_promise', 'teacher_prefix', 'teacher_teach_for_america',
  'teacher_ny_teaching_fellow', 'primary_focus_subject', 'primary_focus_area',
  'secondary_focus_subject', 'secondary_focus_area', 'resource_type', 'poverty_level',
  'grade_level', 'vendor_shipping_charges', 'sales_tax', 'payment_processing_charges',
  'fulfillment_labor_materials', 'total_price_excluding_optional_support',
  'total_price_including_optional_support', 'students_reached', 'total_donations',
  'num_donors', 'eligible_double_your_impact_match', 'eligible_almost_home_match',
  'funding_status', 'date_posted', 'date_completed', 'date_thank_you_packet_mailed',
  'date_expiration'))

essay <- fread("opendata_essays0002/opendata_essays000.csv", encoding = "UTF-8",
  sep = ",", header = FALSE, data.table = FALSE, col.names = c('_projectid',
  '_teacher_acctid', 'title', 'essay'), drop = c(4,5,7,8))
```

5 Data Cleaning and Wrangling

5.1 Changing Data Types

With the fread command, all properties were imported as character strings. Data types were converted to factors, logicals, numbers and dates as needed, as shown below.

```
project$school_latitude <- as.double(project$school_latitude)
project$school_longitude <- as.double(project$school_longitude)

project$school_city <- as.factor(project$school_city)
project$school_state <- as.factor(project$school_state)
project$school_zip <- as.factor(project$school_zip)
project$school_metro <- as.factor(project$school_metro)
project$school_district <- as.factor(project$school_district)
project$school_county <- as.factor(project$school_county)

project$school_charter <- as.logical(toupper(project$school_charter))
project$school_magnet <- as.logical(toupper(project$school_magnet))
project$school_year_round <- as.logical(toupper(project$school_year_round))
project$school_nlns <- as.logical(toupper(project$school_nlns))
project$school_kipp <- as.logical(toupper(project$school_kipp))
project$school_charter_ready_promise <-
  as.logical(toupper(project$school_charter_ready_promise))

project$teacher_prefix <- as.factor(project$teacher_prefix)

project$teacher_teach_for_america <-
  as.logical(toupper(project$teacher_teach_for_america))
project$teacher_ny_teaching_fellow <-
  as.logical(toupper(project$teacher_ny_teaching_fellow))

project$primary_focus_subject <- as.factor(project$primary_focus_subject)
```

```
project$primary_focus_area <- as.factor(project$primary_focus_area)
project$secondary_focus_subject <- as.factor(project$secondary_focus_subject)
project$secondary_focus_area <- as.factor(project$secondary_focus_area)

project$resource_type <- as.factor(project$resource_type)
project$poverty_level <- as.factor(project$poverty_level)
project$grade_level <- as.factor(project$grade_level)

project$vendor_shipping_charges <- as.double(project$vendor_shipping_charges)
project$sales_tax <- as.double(project$sales_tax)
project$payment_processing_charges <- as.double(project$payment_processing_charges)
project$fulfillment_labor_materials <- as.double(project$fulfillment_labor_materials)
project$total_price_excluding_optional_support <-
  as.double(project$total_price_excluding_optional_support)
project$total_price_including_optional_support <-
  as.double(project$total_price_including_optional_support)

project$students_reached <- as.integer(project$students_reached)
project$total_donations <- as.integer(project$total_donations)
project$num_donors <- as.integer(project$num_donors)

project$eligible_double_your_impact_match <-
  as.logical(toupper(project$eligible_double_your_impact_match))
project$eligible_almost_home_match <-
  as.logical(toupper(project$eligible_almost_home_match))

project$funding_status <- as.factor(project$funding_status)

project$date_posted <- as.Date(project$date_posted)
project$date_completed <- as.Date(project$date_completed)
project$date_thank_you_packet_mailed <- as.Date(project$date_thank_you_packet_mailed)
project$date_expiration <- as.Date(project$date_expiration)

project$grade_level <- factor(project$grade_level, levels = c("", "Grades PreK-2",
                                "Grades 3-5", "Grades 6-8", "Grades 9-12"))
project$poverty_level <- factor(project$poverty_level, levels = c("", "low poverty",
                                "moderate poverty", "high poverty", "highest poverty"))
```

5.2 Filtering by Date and Joining the Datasets

The data from DonorsChoose.org covers the period from 2002 until today. Figure 1 illustrates the number of projects posted in each year.

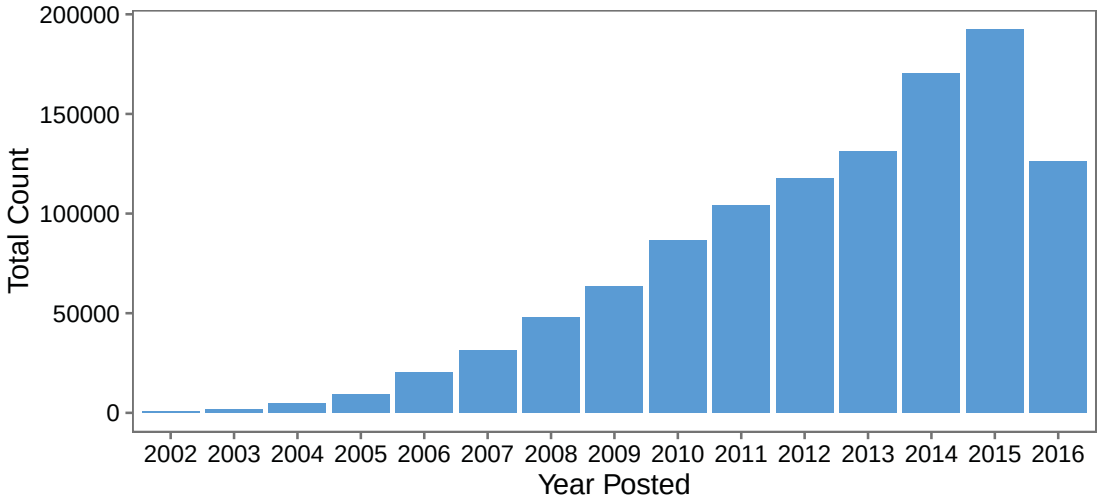

Figure 1. Projects posted by year.

Since the scope of the analysis was limited to 2015 data, the project dataset was filtered by date, and then joined with the essay dataset.

```
project2015 <- project %>%
  subset(date_posted >= "2015-01-01" & date_posted <= "2015-12-31")

project2015 <- project2015 %>%
  left_join(essay, by = c("_projectid", "_teacher_acctid"))
```

5.3 Limiting the Categories for Funding Status

The possible values for funding status include completed, expired, live and reallocated. Since no projects from 2015 are currently active, the "live" status is not seen in our filtered dataset as shown in Table 1.

Table 1. Number of projects at each funding status.

Status	Count
completed	137873
expired	53609
live	0
reallocated	890

The analysis will focus on projects that are either completed (successfully funded) or expired (did not reach funding goal), and the reallocated projects were removed.

```
project2015 <- project2015 %>%
  filter(funding_status == "completed" | funding_status == "expired")
project2015$funding_status <- droplevels(project2015$funding_status)
```

This produced a dataset with 191482 records.

5.4 Dealing with Missing Values

The resource_type property has several missing values, and also a value of "Other" as shown in Table 2.

Table 2. Summary of resource types.

Resource	Count
	4
Books	29728
Other	25025
Supplies	64290
Technology	67386
Trips	4473
Visitors	576

The missing values were assigned the value "Other", since these were assumed to be equivalent.

```
project2015$resource_type[project2015$resource_type == ""] <- "Other"
project2015$resource_type <- droplevels(project2015$resource_type)
```

5.5 Determining Month Posted

The project dataset includes a property for the date the project was posted (date_posted) on DonorsChoose.org. An additional property was created (month_posted) based on the date_posted property.

```
project2015$month_posted <- as.factor(format(project2015$date_posted, '%b'))
project2015$month_posted <- factor(project2015$month_posted, levels = c("Jan",
  "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"))
```

5.6 Focus Area and Subject

The dataset contains values for primary and secondary focus areas and subjects. When donors search for projects through DonorsChoose.org, the website does not differentiate between primary and secondary focus areas. Instead, the project is considered to have two focus areas and two subjects. The data was transformed (as shown in the code below) to have logical properties such as focus_area_applied_learning which will be true if either the primary or secondary focus area is "Applied Learning."

```
checkFocusArea <- function(name) {
  as.logical(project2015$primary_focus_area == name |
               project2015$secondary_focus_area == name)
}

checkFocusSubject <- function(name) {
  as.logical(project2015$primary_focus_subject == name |
               project2015$secondary_focus_subject == name)
}

project2015$focus_area_applied_learning <- checkFocusArea("Applied Learning")
project2015$focus_area_health_sports <- checkFocusArea("Health & Sports")
project2015$focus_area_history_civics <- checkFocusArea("History & Civics")
project2015$focus_area_lit_language <- checkFocusArea("Literacy & Language")
```

```
project2015$focus_area_math_science <- checkFocusArea("Math & Science")
project2015$focus_area_music_arts <- checkFocusArea("Music & The Arts")
project2015$focus_area_special_needs <- checkFocusArea("Special Needs")

project2015$subject_applied_sci <- checkFocusSubject("Applied Sciences")
project2015$subject_character_ed <- checkFocusSubject("Character Education")
project2015$subject_civics_gov <- checkFocusSubject("Civics & Government")
project2015$subject_college_career <- checkFocusSubject("College & Career Prep")
project2015$subject_community <- checkFocusSubject("Community Service")
project2015$subject_early_dev <- checkFocusSubject("Early Development")
project2015$subject_economics <- checkFocusSubject("Economics")
project2015$subject_enviro_sci <- checkFocusSubject("Environmental Science")
project2015$subject_esl <- checkFocusSubject("ESL")
project2015$subject_extracurricular <- checkFocusSubject("Extracurricular")
project2015$subject_financial_lit <- checkFocusSubject("Financial Literacy")
project2015$subject_foreign_lang <- checkFocusSubject("Foreign Languages")
project2015$subject_gym_fitness <- checkFocusSubject("Gym & Fitness")
project2015$subject_health_life_sci <- checkFocusSubject("Health & Life Science")
project2015$subject_health_wellness <- checkFocusSubject("Health & Wellness")
project2015$subject_hist_geog <- checkFocusSubject("History & Geography")
project2015$subject_literacy <- checkFocusSubject("Literacy")
project2015$subject_lit_writing <- checkFocusSubject("Literature & Writing")
project2015$subject_math <- checkFocusSubject("Mathematics")
project2015$subject_music <- checkFocusSubject("Music")
project2015$subject_nutrition <- checkFocusSubject("Nutrition")
project2015$subject_other <- checkFocusSubject("Other")
project2015$subject_parent <- checkFocusSubject("Parent Involvement")
project2015$subject_perform_art <- checkFocusSubject("Performing Arts")
project2015$subject_social_sci <- checkFocusSubject("Social Sciences")
project2015$subject_special_needs <- checkFocusSubject("Special Needs")
project2015$subject_team_sports <- checkFocusSubject("Team Sports")
project2015$subject_visual_arts <- checkFocusSubject("Visual Arts")
```

6 Exploratory Data Analysis

6.1 Relationships between Project Properties and Funding Status

The relationships between funding status and various properties of the classroom projects were investigated. Highlights from this exploratory data analysis are summarized in this section.

Every project posted on DonorsChoose.org is categorized by poverty level, which refers to the percentage of students at a given school who qualify for free or reduced-price lunch, and is considered a measure of economic need. Schools with 65% or more of its students receiving these lunches are denoted as "highest poverty." Figure 2 shows the number of projects at each poverty level, and the proportion of those projects that were successfully funded. From the plots, it can be seen that projects from highest poverty schools have a slightly greater success rate (73.3%) compared to other schools (low poverty - 71.7%, moderate - 71.5%, high - 69.6%), and that the majority of projects fall under the "highest poverty" category.

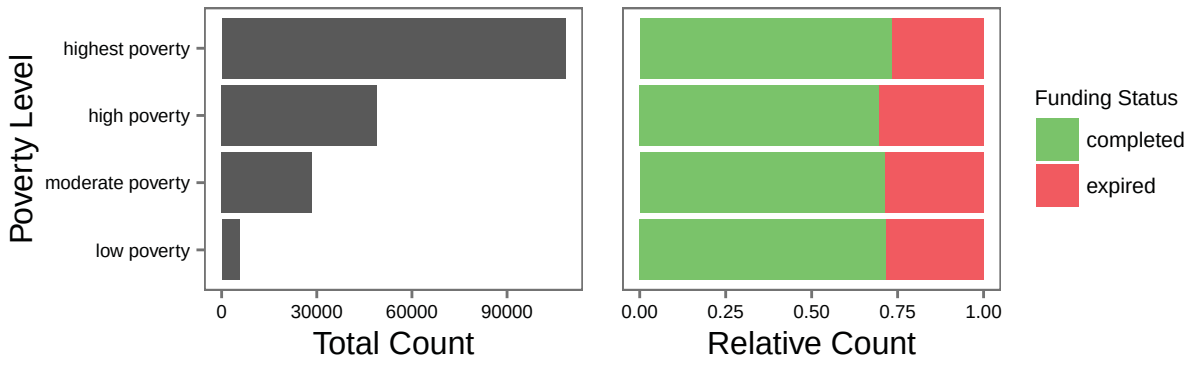

Figure 2. a) Number of projects posted in 2015 at each poverty level. b) Proportion of those projects successfully funded.

Projects are also categorized according to metro region (urban, suburban or rural). Figure 3 shows the number of projects associated with each metro region, and the proportion of those projects that were successfully funded. Schools in urban areas have a higher success rate (74.4%) compared with schools in either suburban (70.7%) or rural (66.8%) regions.

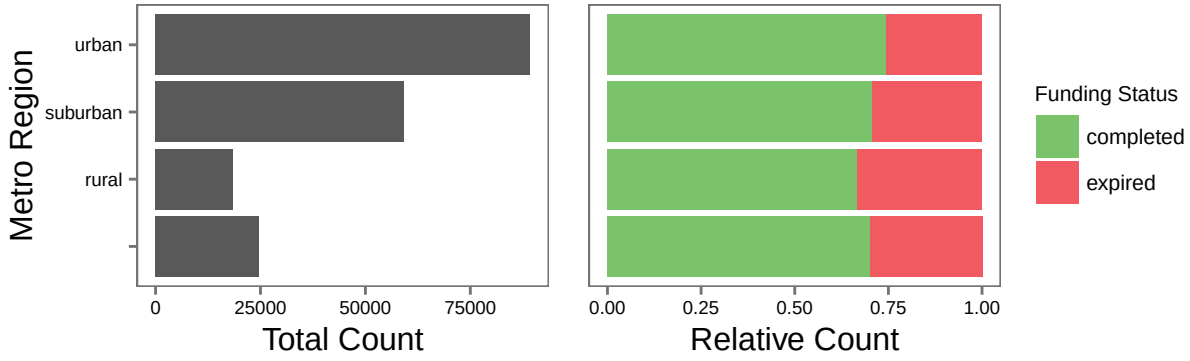

Figure 3. a) Number of projects posted in 2015 in each metro region. b) Proportion of those projects successfully funded.

Figure 4 shows the number of projects associated with each type of resource requested, and the proportion of those projects that were successfully funded. Trip-related projects had the highest success rate of 79.3%, and technology-based projects had the lowest rate of success (65.0%).

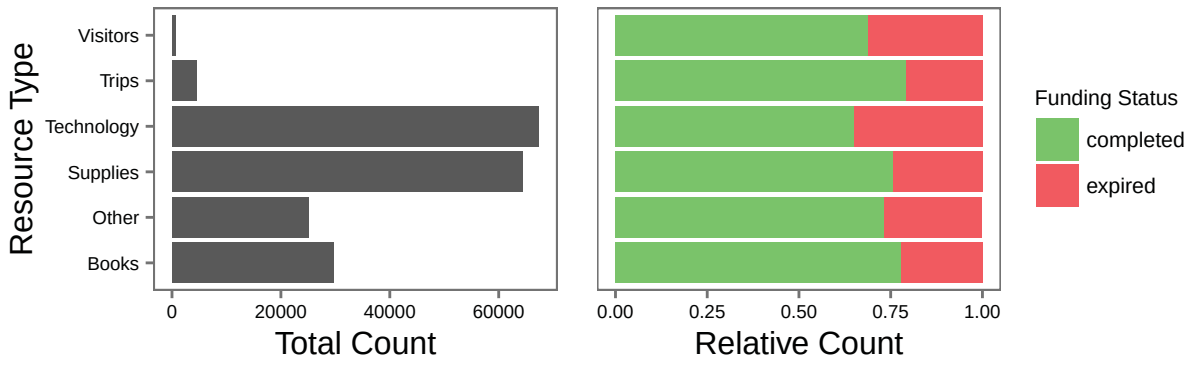

Figure 4. a) Number of projects posted in 2015 for each type of resource requested. b) Proportion of those projects successfully funded.

The number of projects associated with each primary focus subject is shown in Figure 5, together with the proportion of those projects that were successfully funded. A large number of projects were submitted under the categories of Mathematics, Literature & Writing, and Literacy. The success rate ranges from 64.9% for Gym & Fitness projects to 80.5% for Nutrition projects.

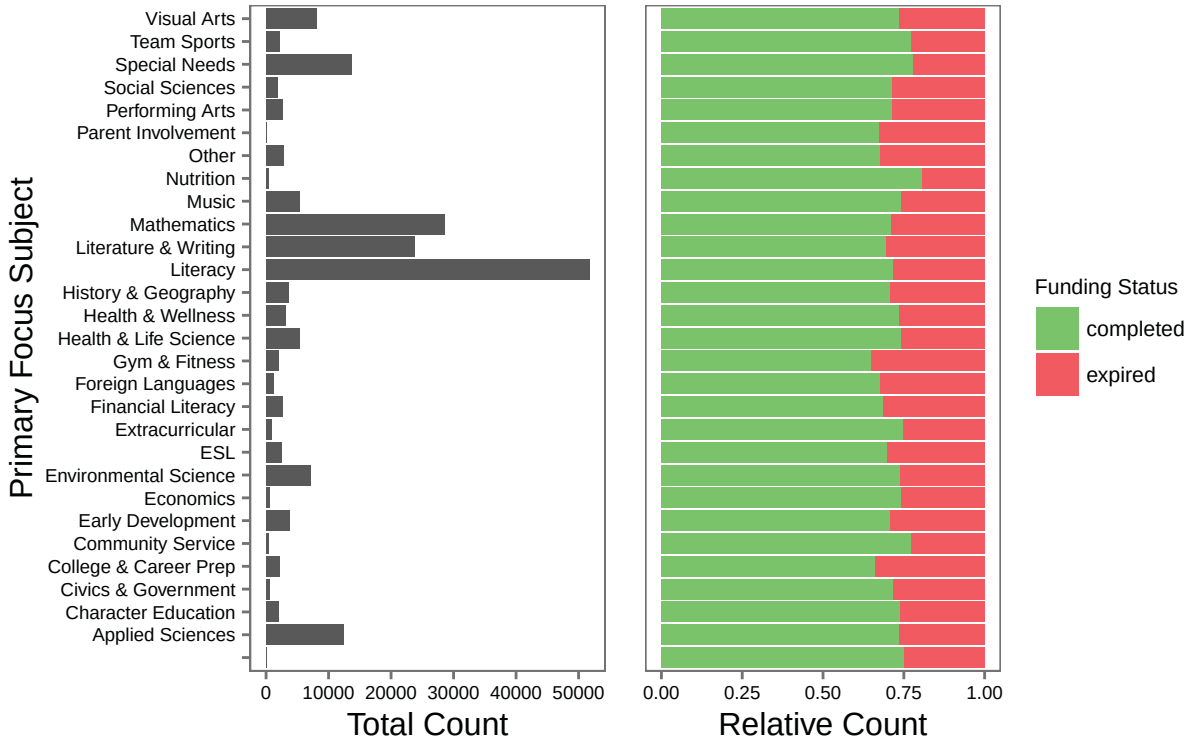

Figure 5. a) Number of projects posted in 2015 for each primary focus subject. b) Proportion of those projects successfully funded.

Figure 6 shows the number of projects posted in each month of 2015, and the success rate by month. A large number of projects was posted in September (typically the start of the school year), and the highest success rate was seen with projects posted in December (83.2%). The month of May had the lowest success rate of 59.8%.

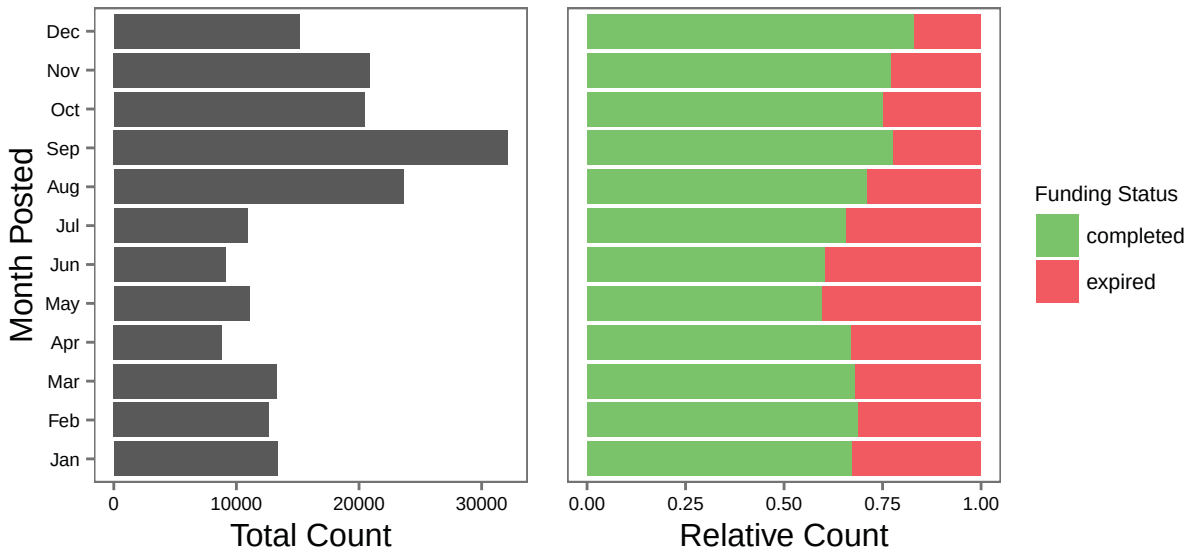

Figure 6. a) Number of projects posted in each month of 2015. b) Proportion of those projects successfully funded.

Each project has two associated price values:

- Total price excluding optional support - the cost of all materials, taxes and fees needed for a project to be successfully funded.
- Total price including optional support - the total price, plus optional funds that go directly to DonorsChoose.org.

Since the "total price excluding optional support" more closely relates to a project's success, it was considered for further analysis. Figure 7 shows that a lower total price is associated with a higher success rate.

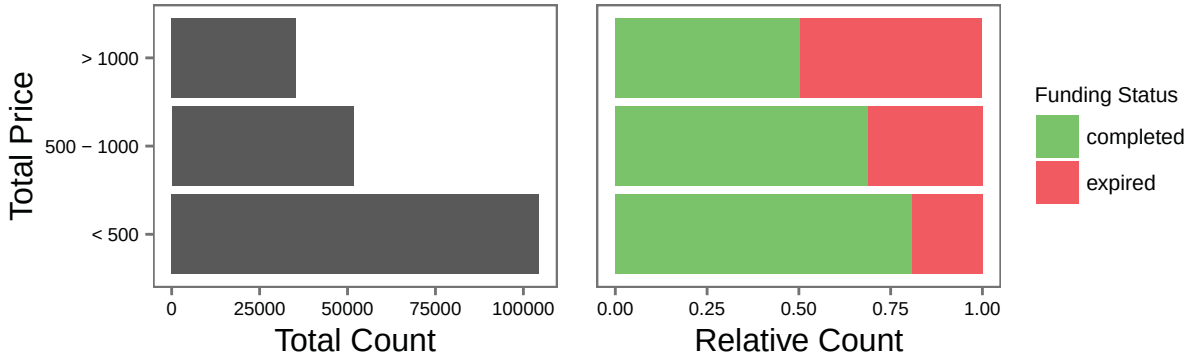

Figure 7. a) Number of projects posted in 2015 by total price (excluding optional support). b) Proportion of those projects successfully funded.

6.2 Geographic Analysis

The choroplethr and choroplethrMaps packages were used to construct maps to illustrate the percentage of projects successfully completed in each state and county across the US in 2015. The maps (Figures 8 and 9) clearly show that there are large differences in project funding across the country, with some counties showing success rates less than 50%.

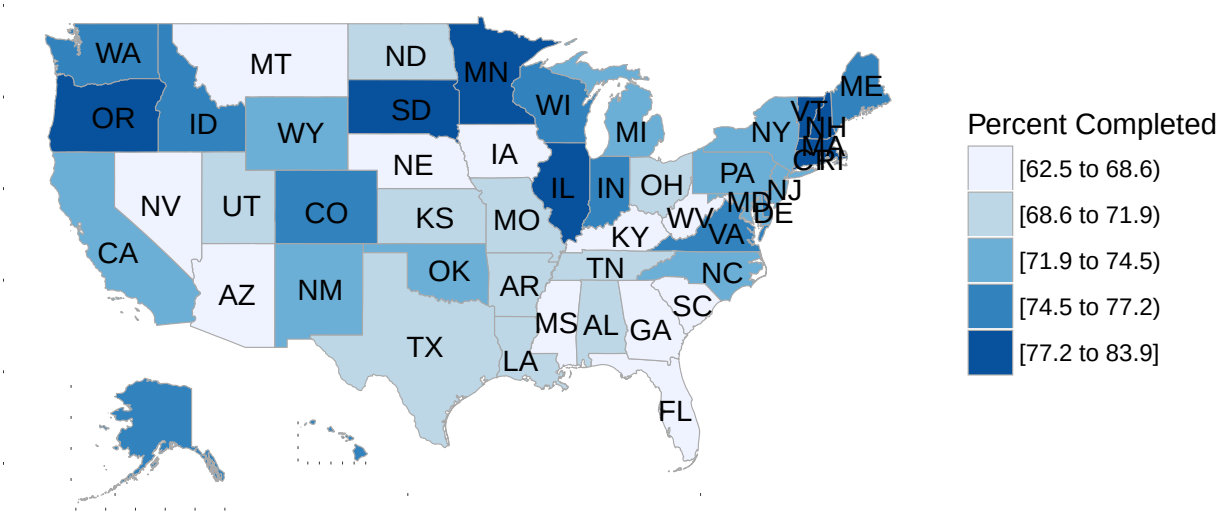

Figure 8. Percentage of projects completed in each state in 2015.

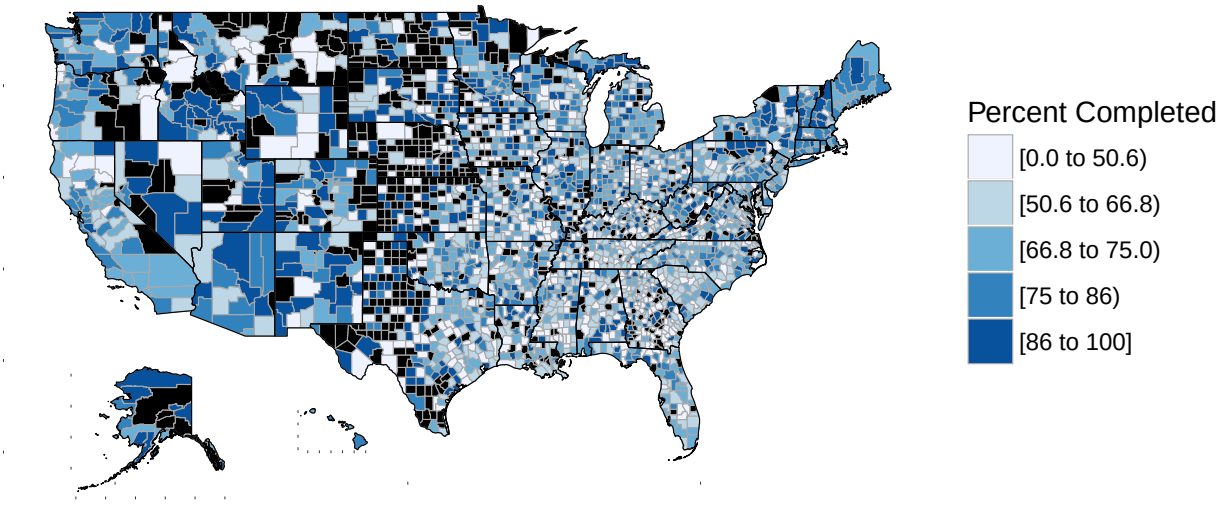

Figure 9. Percentage of projects completed in each county in 2015. Counties shaded in black did not post any projects in 2015.

6.3 Sentiment Analysis

6.3.1 Calculation of Sentiment Scores

Sentiment analysis was conducted using two lexicons from the tidytext package, specifically the sentiment lexicon from Bing Liu and collaborators and the lexicon of Finn Arup Nielsen (AFINN).

The AFINN lexicon provides a positivity score for each word, from -5 (most negative) to +5 (most positive). Positivity scores were calculated for the words in each essay, and then averaged to get an overall sentiment score for the essay.

The Bing lexicon rates words as positive or negative. The total number of positive terms and total number of negative terms were determined for each essay. The difference was calculated and then divided by the essay length to get another sentiment score.

Due to the extensive size of the essay portion of the dataset and limited memory on the workstation used for this analysis, the sentiment calculations (shown below) were run in batches of 10,000 records.

```
AFINN <- subset(sentiments, lexicon == "AFINN", select = c(word, score))
bing <- get_sentiments("bing")

sentiment_data <- data.frame()
bing_data <- data.frame()

for (i in 0:19) {
  firstrow <- 1 + i * 10000
  lastrow <- (i + 1) * 10000

  essay_subset <- project2015[firstrow:lastrow,] %>%
    select(`_projectid`, `_teacher_acctid`, essay)

  essay_words <- essay_subset %>%
    unnest_tokens(word, essay)

  essay_word_count <- essay_words %>%
    count(`_projectid`, `_teacher_acctid`) %>%
    rename(wordcount = n)

  essay_words <- essay_words %>%
    filter(!word %in% stop_words$word, str_detect(word, "^[a-z']+$"))

  afinn_words <- essay_words %>%
    inner_join(AFINN, by = "word")

  afinn_sentiment <- afinn_words %>%
    group_by(`_projectid`, `_teacher_acctid`) %>%
    summarize(afinn_score = mean(score))

  bing_words <- essay_words %>%
    inner_join(bing, by = "word")

  bing_sentiment <- bing_words %>%
    count(`_projectid`, `_teacher_acctid`, sentiment) %>%
    spread(sentiment, n, fill = 0) %>%
    mutate(bing_count = positive - negative)

  essay_sentiment <- essay_word_count %>%
    left_join(afinn_sentiment, by = c("_projectid", "_teacher_acctid")) %>%
    left_join(bing_sentiment, by = c("_projectid", "_teacher_acctid")) %>%
    mutate(bing_score = 100*bing_count/wordcount)

  sentiment_data <- sentiment_data %>% bind_rows(essay_sentiment)
  bing_data <- bing_data %>% bind_rows(bing_words)
}
```

```r
project2015 <- project2015 %>%
  left_join(sentiment_data, by = c("_projectid","_teacher_acctid"))

# assume that missing wordcount means that the essay contained no words
project2015$wordcount[is.na(project2015$wordcount)] <- 0

# assume that missing afinn_score means the essay contained no words in the
# AFINN lexicon and therefore should have a neutral sentiment score of zero
project2015$afinn_score[is.na(project2015$afinn_score)] <- 0


# assume that missing bing_score means the essay contained no words in the
# Bing lexicon and therefore should have a neutral sentiment score of zero
project2015$bing_score[is.na(project2015$bing_score)] <- 0
```

6.3.2 Distribution of Sentiment Scores

Figures 10 and 11 show the distribution of AFINN and Bing sentiment scores for the essays from 2015. Both plots indicate that the majority of essays take on a positive tone (score greater than zero).

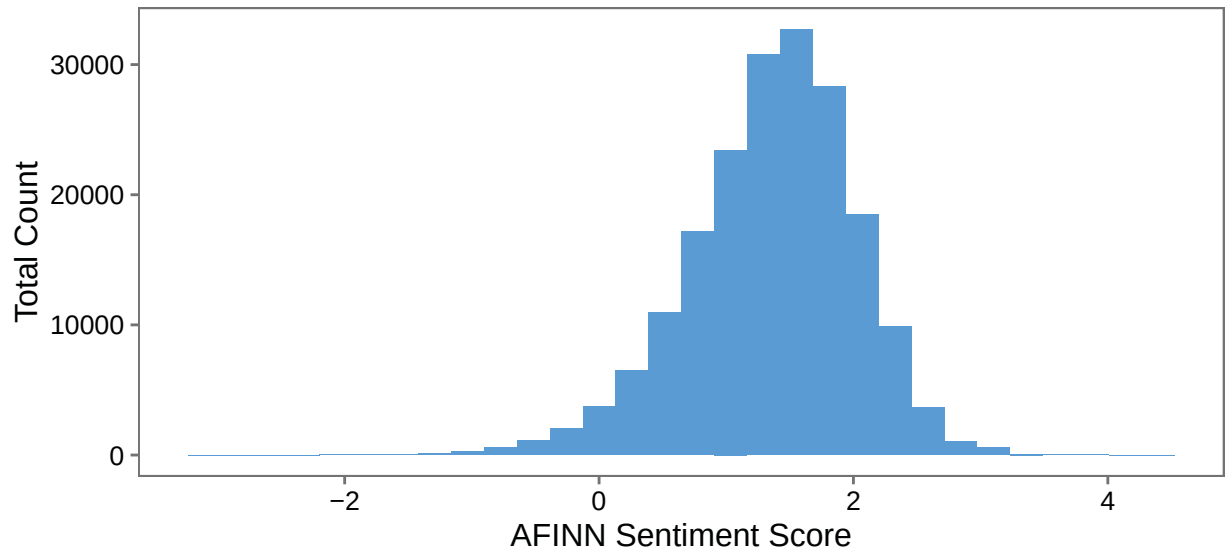

Figure 10. Distribution of AFINN Sentiment Scores for essays posted in 2015.

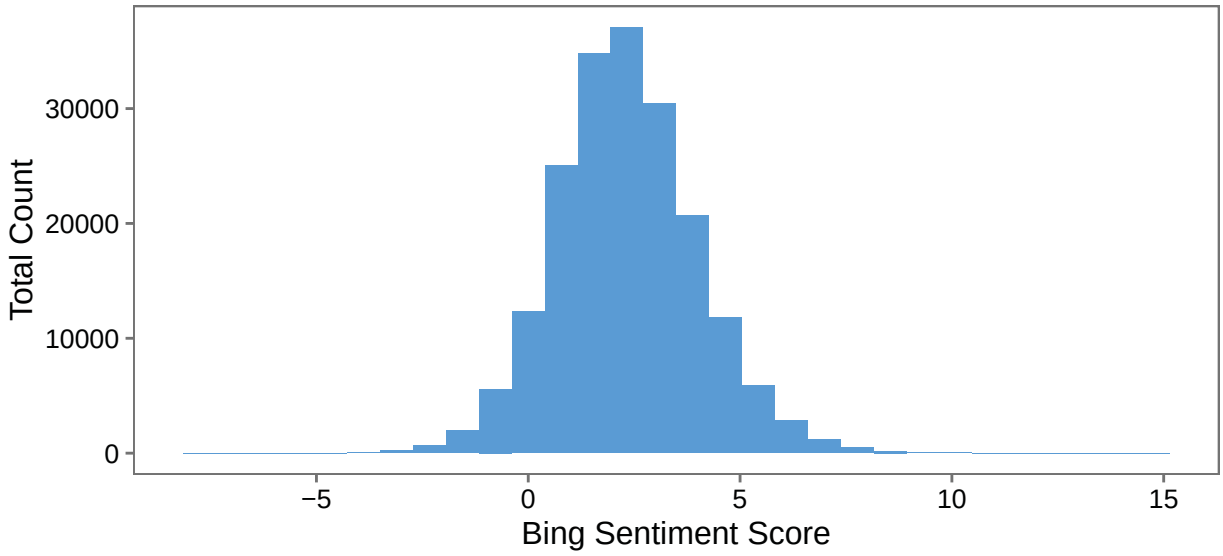

Figure 11. Distribution of Bing Sentiment Scores for essays posted in 2015.

6.3.3 Text of Positive and Negative Essays

According to the Bing sentiment scores calculated above, the most positive essay (highest score of 14.6) is from a project titled "Comfy Carpet for Reading!". The essay text is shown below, and takes on a very positive tone:

> *Picture yourself curled up with a good book sitting in your favorite comfy spot. Doesn't that sound nice? I would love for my little learners to have that opportunity too so they can really enjoy reading! I am so blessed to teach 1st grade at a wonderful and fun school in Las Vegas. Our school is high performing and we are also a 5-star gold school the highest honor awarded to any school in Nevada. My class is filled with the most friendly lovable respectful kids you will ever meet! My students need a comfy cozy carpet so they will have a comfortable place to sit while reading. Being comfortable will encourage them to use our classroom library and explore books so their reading time will be more enjoyable. I believe offering a fun and cozy place to sit will also help them to develop the skill of reading for pleasure– one of the most important keys to becoming excellent readers. Your help will enable young students to develop a love of reading and become lifelong readers. Your donations will provide a comfy cozy library corner for many appreciative kids. Best of all this carpet will benefit students in my classroom for years to come. Thank you for helping my students get Comfy Cozy and become better readers!*

The essay taking on the most negative tone (lowest Bing score of -7.9) is from a project titled "Help with Anxiety". The essay text is shown below:

> *A world were they feel nervous and scared. Lots of children are scared or nervous. Some of my students get nervous around loud noises or act out because they can't communicate. I am trying to help them deal with their anxiety. My students are physically challenged and developmentally delayed. Some have trouble talking because they do not have the vocal cord strength or anxiety prevents them from talking like a mainstream student does. I try to use*

different modalities to help the students deal with their anxiety so they do not act out. The pressure jacket is something that is very helpful to a child with anxiety issues. The pressure helps them feel safe. The jacket is something that can be used at high stress times and for students who have trouble with transitions. I am also requesting multi-sensory toys(sensory and puffer balls). These object can help students with ADHD and anxiety issues. The jacket and toys can help a child keep calm in a world that is stressful for them. The pressure jacket and sensory toys will make a difference in my special students lives. It will help them more easily participate in a world that scares them. It will allow them to feel safe in a world that causes them anxiety.

6.3.4 Most Common Positive and Negative Terms

Using the Bing lexicon, the 50 most commonly used positive and negative terms were identified, as shown in the plot in Figure 12. The most common positive term is "love", and the most common negative term is "hard." The word cloud shown in Figure 13 illustrates some of the more commonly used positive and negative terms.

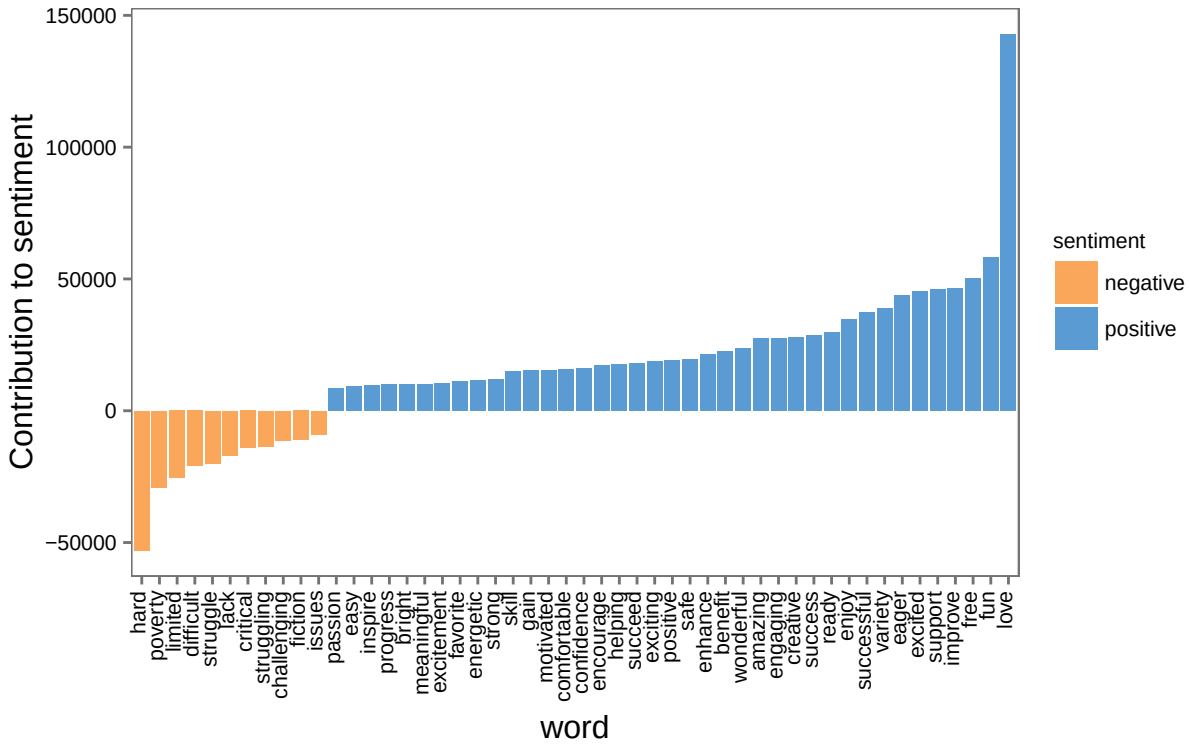

Figure 12. Number of occurrences of the 50 most common positive and negative terms from essays posted in 2015.

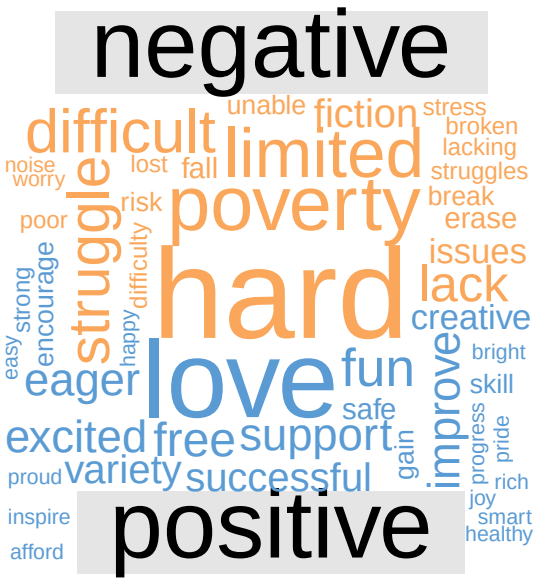

Figure 13. Word cloud visualization of common positive and negative terms from essays posted in 2015.

7 Model Building

7.1 Overview

The objective of this capstone project is to provide guidance to teachers posting projects on DonorsChoose.org. A classification tree was constructed to predict a project's "funding_status" (completed or expired) in order to gain insight into what properties of a project improve its chances of success. The CART method was chosen, since it is considered to be more interpretable compared to other methods such as Random Forest.

Only properties of the project and essay that are within the teacher's control were considered when building the model. For example, a teacher would have no control over school location, poverty level or grade level, but would be able to change the project focus, resources requested and essay content.

The independent variables were chosen to be:

- resource_type
- total_price_excluding_optional_support
- focus_area (7 properties)
- subject (28 properties)
- wordcount
- afinn_score
- bing_score
- month_posted

The dependent variable was:

- funding_status

7.2 Training and Test Set

The dataset from 2015 was split into training and test sets, with 70% of the records assigned to the training set.

```
project2015$completed <- as.logical(project2015$funding_status == "completed")

set.seed(1000)
split <- sample.split(project2015$completed, SplitRatio = 0.7)

train <- subset(project2015, split == TRUE)
test <- subset(project2015, split == FALSE)
```

This produced a training set with 134037 records, and a test set with 57445 records.

7.3 Cross Validation

Five-fold cross validation was performed to optimize the value of the complexity parameter (cp), as shown below.

```
fitControl <- trainControl(method = "cv", number = 5)
cartGrid <- expand.grid(.cp=(1:10)*0.0001)

cartCV <- train(funding_status ~ resource_type + total_price_excluding_optional_support +
  focus_area_applied_learning + focus_area_health_sports + focus_area_history_civics +
  focus_area_lit_language + focus_area_math_science + focus_area_music_arts +
  focus_area_special_needs + subject_applied_sci + subject_character_ed +
  subject_civics_gov + subject_college_career + subject_community + subject_early_dev +
  subject_economics + subject_enviro_sci + subject_esl + subject_extracurricular +
  subject_financial_lit + subject_foreign_lang + subject_gym_fitness +
  subject_health_life_sci + subject_health_wellness + subject_hist_geog +
  subject_literacy + subject_lit_writing + subject_math + subject_music +
  subject_nutrition + subject_other + subject_parent + subject_perform_art +
  subject_social_sci + subject_special_needs + subject_team_sports + subject_visual_arts +
  wordcount + afinn_score + bing_score + month_posted, data = train, method = "rpart",
  trControl = fitControl, tuneGrid = cartGrid)
```

Results from the cross-validation are shown below, where the optimum cp value is 0.0002, since it is associated with the highest accuracy.

```
CART

134037 samples
    41 predictor
     2 classes: 'completed', 'expired'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 107229, 107230, 107229, 107230, 107230
Resampling results across tuning parameters:

  cp     Accuracy   Kappa
  1e-04  0.7403627  0.2237351
  2e-04  0.7444138  0.2231421
  3e-04  0.7441751  0.2217445
  4e-04  0.7438990  0.2251397
  5e-04  0.7437424  0.2254296
  6e-04  0.7434663  0.2271645
  7e-04  0.7425412  0.2229798
  8e-04  0.7412953  0.2049295
  9e-04  0.7410864  0.2006783
  1e-03  0.7410267  0.2010635

Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was cp = 2e-04.
```

7.4 Building CART Model and Pruning the Tree

A CART model was built using the optimized cp parameter with the following R code.

```
modelcart <- rpart(funding_status ~ resource_type +
  total_price_excluding_optional_support + focus_area_applied_learning +
  focus_area_health_sports + focus_area_history_civics + focus_area_lit_language +
  focus_area_math_science + focus_area_music_arts + focus_area_special_needs +
  subject_applied_sci + subject_character_ed + subject_civics_gov +
  subject_college_career + subject_community + subject_early_dev + subject_economics +
  subject_enviro_sci + subject_esl + subject_extracurricular + subject_financial_lit +
  subject_foreign_lang + subject_gym_fitness + subject_health_life_sci +
  subject_health_wellness + subject_hist_geog + subject_literacy + subject_lit_writing +
  subject_math + subject_music + subject_nutrition + subject_other + subject_parent +
  subject_perform_art + subject_social_sci + subject_special_needs + subject_team_sports +
  subject_visual_arts + wordcount + afinn_score + bing_score + month_posted,
  data = train, method = "class", cp = 0.0002)
```

The fitted tree's CP table (Table 3) provides a summary of the model's overall fit. The table lists the trees associated with the model in order from smallest tree (no splits) to largest tree (72 splits).

Table 3. CP table for the CART model.

CP	nsplit	rel error	xerror	xstd
0.04434	0	1	1	0.00438
0.01572	1	0.9557	0.9558	0.004319
0.008927	2	0.9399	0.9401	0.004296
0.003318	3	0.931	0.9315	0.004284
0.001612	5	0.9244	0.9254	0.004274
0.001372	7	0.9211	0.9227	0.00427
0.001352	9	0.9184	0.9216	0.004269
0.001173	14	0.9116	0.9177	0.004263
0.000906	15	0.9104	0.9145	0.004258
0.0008261	16	0.9095	0.9118	0.004254
0.0007461	18	0.9078	0.9117	0.004254
0.0005996	19	0.9071	0.9114	0.004253
0.0005063	21	0.9059	0.9092	0.00425
0.0004797	24	0.904	0.9086	0.004249
0.000453	26	0.9031	0.9083	0.004248
0.0004264	33	0.8992	0.908	0.004248
0.0003464	34	0.8988	0.9077	0.004248
0.0002931	36	0.8981	0.9071	0.004247
0.0002825	38	0.8975	0.9068	0.004246
0.0002754	46	0.8943	0.9069	0.004246
0.0002665	50	0.893	0.9066	0.004246
0.0002532	52	0.8925	0.9061	0.004245
0.0002398	59	0.8906	0.9062	0.004245
0.000231	61	0.8901	0.9059	0.004245
0.0002132	65	0.8891	0.9064	0.004246
2e-04	72	0.8876	0.9081	0.004248

The 1-SE rule can be used to find the best number of splits. This method takes the smallest cross validation error (xerror) and adds the corresponding standard error (xstd). It then finds the simplest tree with a cross-validated error that is less than this total. In this case, the minimum xerror is 0.9068 with a corresponding xstd of 0.004246 at the tree with 38 splits. The sum of these two values is 0.911046. The simplest tree with a cross-validated error less than this sum is the tree with an xerror of 0.9092 (and 21 splits).

The model was pruned with the 21 split tree as the target. This was accomplished using a cp value in the range of $0.0005063 < \text{cp} < 0.0005996$. In this case, 0.00055 was chosen. The pruned tree is shown in Figure 14. Additional details for each node of the tree can be found in Appendix 1.

```
prunedcart <- prune(modelcart, cp = 0.00055)
```

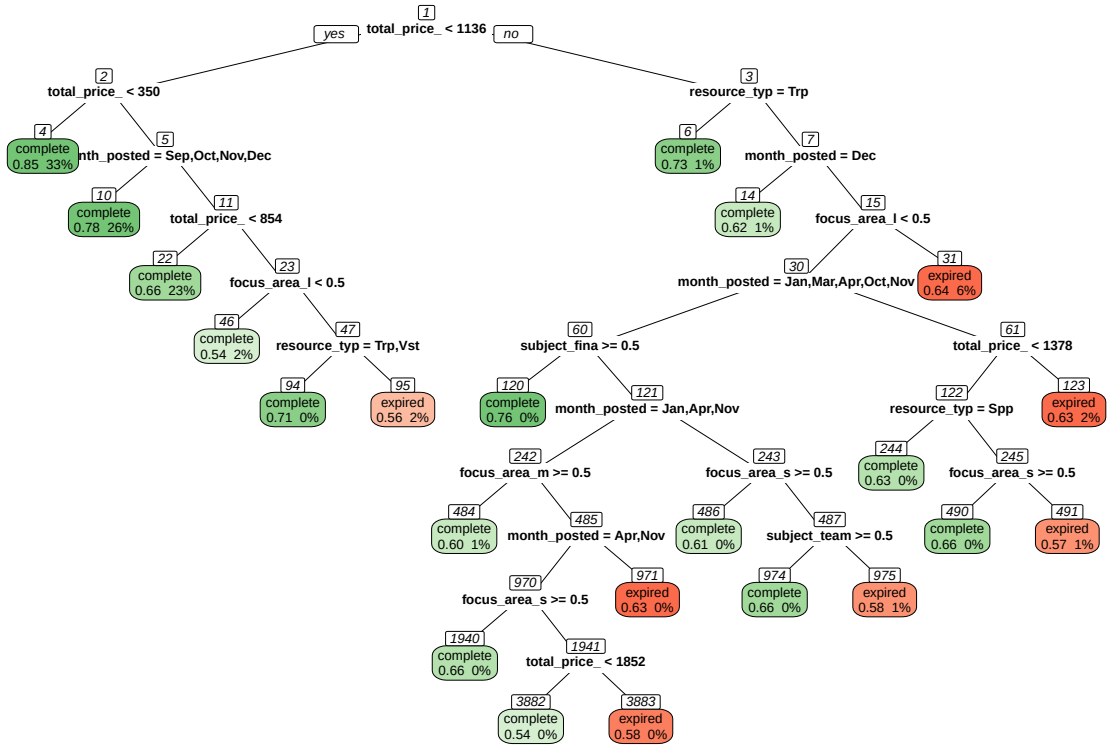

Figure 14. Pruned classification tree. Leaf node labels include the name of the fitted class (complete or expired), the probability of the fitted class, and the percentage of observations in the node. Leaf nodes are colored based on the fitted class: green for complete and red for expired, with a more intense color indicating greater probability.

7.5 Evaluating the Model's Fit

The confusion matrix for the training set is shown below.

Table 4. Confusion matrix for the training set. Row labels are the actual funding status options. Column labels are the predicted funding status options.

	completed	expired
completed	90609	5902
expired	28092	9434

From the confusion matrix, the accuracy of the model on the training set is 74.6%. The baseline model, where all projects are predicted to belong to the majority class (funding_status = completed), has an accuracy of 72.0%. The CART model is a modest improvement over the baseline model.

The confusion matrix for the test set is shown in Table 5.

Table 5. Confusion matrix for the test set. Row labels are the actual funding status options. Column labels are the predicted funding status options.

	completed	expired
completed	38782	2580
expired	12043	4040

The accuracy of the model on the test set is 74.5%, which is similar to what was observed for the training set. This indicates that the pruned tree is not overfitting the dataset.

The CART model can also be evaluated using a ROC (Receiver Operating Characteristics) curve, which plots the true positive rate against the false positive rate. The class probabilities were calculated for each observation in the training set as shown in the following code. The ROC curve was generated using class probability as a cutoff (Figure 15).

```
predTrain <- predict(prunedcart, type = "prob")[,2]
predictionTrain = prediction(predTrain, train$funding_status)
trainAUC <- as.numeric(performance(predictionTrain, "auc")@y.values)

plot(performance(predictionTrain, "tpr", "fpr"))
abline(0, 1, lty = 2)
```

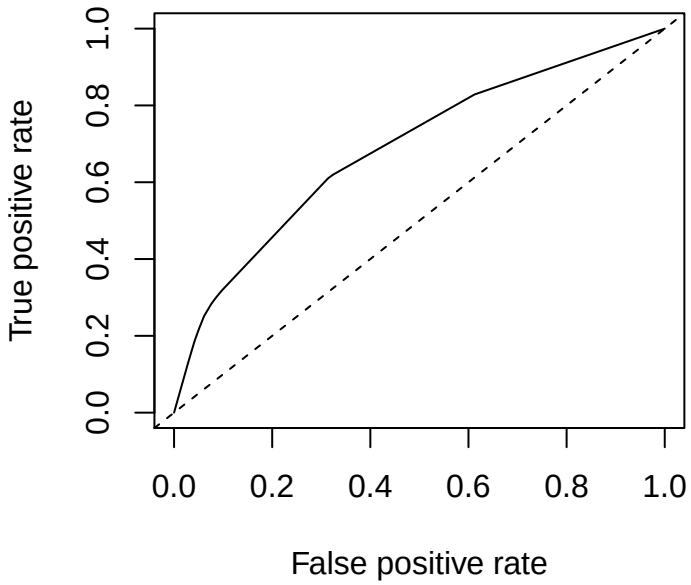

Figure 15. ROC curve for the training set.

The AUC (area under the curve) associated with this ROC plot is 0.69. The same procedure was applied to the test set, and the corresponding ROC plot is shown in Figure 16.

```
predTest <- predict(prunedcart, newdata = test, type = "prob")[,2]
predictionTest = prediction(predTest, test$funding_status)
testAUC <- as.numeric(performance(predictionTest, "auc")@y.values)

plot(performance(predictionTest, "tpr", "fpr"))
abline(0, 1, lty = 2)
```

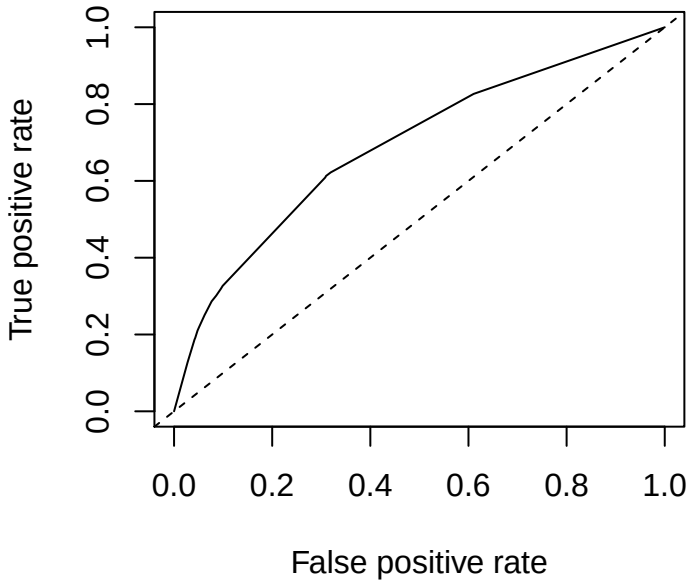

Figure 16. ROC curve for the training set.

The test set AUC is 0.69, which is similar to the AUC of the training set.

7.6 Interpreting the CART Tree

Although the CART model developed here provides only a modest improvement over the baseline model (accuracy of 74.6% vs 72.0%), and did not generate a large AUC value, there is still valuable information to be gained from the CART tree.

The leaf nodes were examined, and those labelled as "complete" with high class probabilities were considered for further analysis. Node 4 (the left-most leaf node) in Figure 14 corresponds to projects with total price less than $350.31. The class probability associated with this node is 0.85 (or 85%). We can infer that any project costing less than $350.31 will improve its chances of success to 85% (instead of the average success rate of 72%).

This statement was evaluated using the test set. The table below shows the funding status for test set projects costing less than $350.31. From this table we can see that 85% of these test set projects were successfully funded, which matches the success rate seen for the training set. Therefore, we can assume this is a good estimate of project success.

Table 6. Funding status for test set projects posted in 2015 with total price (excluding opt. support) less than $350.31.

Status	Count
completed	16108
expired	2790

Leaf node 10 (second from the left) in Figure 14 represents projects posted from Sept. to Dec. with total price between $350.31 and $1136.50. The class probability associated with this node is 0.78 (or 78%). We can infer that any project posted from Sept to Dec, and costing $350.31 to $1136.50 will improve its chances of success to 78%.

This statement was evaluated using the test set. The table below shows the funding status for test set projects falling in this month and price range. From the data in this table we can see that 79% of these test set projects were successfully funded, which is similar to the success rate seen for the training set. Therefore, we can assume this is a good estimate of project success.

Table 7. Funding status for test set projects posted from September to December 2015 with total price (excluding opt. support) between $350.31 and $1136.50.

Status	Count
completed	11975
expired	3268

Leaf node 6 in Figure 14 corresponds to projects costing more than $1136.50 where the resource type is "Trips". The class probability associated with this node is 0.73 (or 73%). We can infer that any classroom project involving a trip that costs more than $1136.50 has a probability of success of 73%.

This statement was evaluated using the test set. The table below shows the funding status for test set projects involving a trip costing more than $1136.50. According to this table, 73% of these test set projects were successfully funded, which is the same as the success rate seen for the training set. Therefore, we can assume this is a good estimate of project success.

Table 8. Funding status for test set projects involving class trips from 2015 with total price (excluding opt. support) greater than $1136.50.

Status	Count
completed	395
expired	148

The only other leaf node labelled as "complete" with above average class probabilities (> 0.72) is node 120. However, this leaf node represents a very small percentage of the dataset (0.2%), and was therefore not considered for further analysis.

8 Discussion

From the exploratory data analysis, we can see that a number of different factors impact a project's chances of success. Several of these factors are not within a teacher's control, including school location and poverty level.

Projects from "highest poverty" schools have a slightly better success rate than lower poverty schools. The

DonorsChoose.org website allows donors to filter projects by the "highest poverty"" category, but does not allow filtering by other poverty levels, which may explain why these projects tend to be more successful.

School location has a major impact on project success rate. From the geographic analysis, we can see that certain states and counties have much higher success rates. And the comparison of metro regions revealed that schools in urban areas have a better success rate. There may be a connection between metro region and poverty level, which could be explored in future studies.

The main focus of this capstone project was to examine the factors that are within a teacher's control, including project topic, resources requested, total cost and essay content. The classification tree constructed using these variables revealed that projects costing less than $350 improved their chances of success to 85%.

Projects costing $350 to $1136 had a success rate of 78% provided that they were posted between September and December. Donors may be in more of a giving mood around the holidays, which may explain the improved success rate at this time of year.

Projects costing more than $1136 generally had lower success rates, with the exception of trip-based projects which still came in slightly above average at 73%. This is in agreement with results seen during exploratory data analysis, where trip-based projects had the highest rate of success.

9 Conclusion

Using 2015 data from DonorsChoose.org, a classification tree was built to predict whether a project will receive full funding based on a number of different factors that are within a teacher's control. Total price and month posted had the largest impact on improving a project's chance of success. Based on the analysis of the classification tree, several recommendations are proposed for teachers using DonorsChoose.org:

1. For best chances of success, keep the project costs under $350.
2. If the project will cost more ($350 - $1136), then be sure to post the project early in the school year (September to December).
3. Avoid letting the costs get above $1136, unless considering a trip-based project.

10 Future Work

This capstone project focused on the project and essay datasets from DonorsChoose.org. Additional insight can likely be gained from the other datasets available on the website, including the resource and donations datasets.

Suggestions for future work include:

- Using the resources dataset (which breaks down project costs into individual items) to determine if average item price and total number of items requested have an impact on funding success. Do donors prefer contributing to one big cost item, or do they prefer having their money go to a large number of inexpensive items?
- Using the donations dataset, the distance between each donor's location and the school could be calculated to see if donors typically donate to nearby schools. This information may help teachers understand where to focus their fundraising efforts.
- Using the donations dataset, the characteristics of repeat donors vs one-time donors could be analyzed. This information may help teachers understand how to get donors to return and fund future classroom projects.

Appendix 1

Details for each node of the pruned CART tree from Figure 14 are shown below.

```
n= 134037

node), split, n, loss, yval, (yprob)
      * denotes terminal node

   1) root 134037 37500 completed (0.720 0.280)
   2) total_price_excluding_optional_support< 1.14e+03 115485 27400 completed (0.763 0.237)
   4) total_price_excluding_optional_support< 350 43847  6440 completed (0.853 0.147) *
   5) total_price_excluding_optional_support>=350 71638 21000 completed (0.707 0.293)
  10) month_posted=Sep,Oct,Nov,Dec 35459  7790 completed (0.780 0.220) *
  11) month_posted=Jan,Feb,Mar,Apr,May,Jun,Jul,Aug 36179 13200 completed (0.635 0.365)
  22) total_price_excluding_optional_support< 854 31410 10800 completed (0.656 0.344) *
  23) total_price_excluding_optional_support>=854 4769  2380 completed (0.501 0.499)
  46) focus_area_lit_language< 0.5 2334  1060 completed (0.545 0.455) *
  47) focus_area_lit_language>=0.5 2435  1120 expired (0.458 0.542)
  94) resource_type=Trips,Visitors 130    38 completed (0.708 0.292) *
  95) resource_type=Books,Other,Supplies,Technology 2305  1020 expired (0.444 0.556) *
   3) total_price_excluding_optional_support>=1.14e+03 18552  8440 expired (0.455 0.545)
   6) resource_type=Trips 1256   333 completed (0.735 0.265) *
   7) resource_type=Books,Other,Supplies,Technology,Visitors 17296  7520 expired (0.435 0.565)
  14) month_posted=Dec 1349   507 completed (0.624 0.376) *
  15) month_posted=Jan,Feb,Mar,Apr,May,Jun,Jul,Aug,Sep,Oct,Nov 15947  6680 expired (0.419 0.581)
  30) focus_area_lit_language< 0.5 8455  3970 expired (0.469 0.531)
  60) month_posted=Jan,Mar,Apr,Oct,Nov 3955  1850 completed (0.531 0.469)
 120) subject_financial_lit>=0.5 204    48 completed (0.765 0.235) *
 121) subject_financial_lit< 0.5 3751  1800 completed (0.519 0.481)
 242) month_posted=Jan,Apr,Nov 2354  1060 completed (0.552 0.448)
 484) focus_area_math_science>=0.5 1381   553 completed (0.600 0.400) *
 485) focus_area_math_science< 0.5 973   471 expired (0.484 0.516)
 970) month_posted=Apr,Nov 738   355 completed (0.519 0.481)
1940) focus_area_special_needs>=0.5 145    50 completed (0.655 0.345) *
1941) focus_area_special_needs< 0.5 593   288 expired (0.486 0.514)
3882) total_price_excluding_optional_support< 1.85e+03 322   147 completed (0.543 0.457) *
3883) total_price_excluding_optional_support>=1.85e+03 271   113 expired (0.417 0.583) *
 971) month_posted=Jan 235    88 expired (0.374 0.626) *
 243) month_posted=Mar,Oct 1397   647 expired (0.463 0.537)
 486) focus_area_special_needs>=0.5 200    78 completed (0.610 0.390) *
 487) focus_area_special_needs< 0.5 1197   525 expired (0.439 0.561)
 974) subject_team_sports>=0.5 95    32 completed (0.663 0.337) *
 975) subject_team_sports< 0.5 1102   462 expired (0.419 0.581) *
  61) month_posted=Feb,May,Jun,Jul,Aug,Sep 4500  1860 expired (0.414 0.586)
 122) total_price_excluding_optional_support< 1.38e+03 1256   601 completed (0.521 0.479)
 244) resource_type=Supplies 463   171 completed (0.631 0.369) *
 245) resource_type=Books,Other,Technology,Visitors 793   363 expired (0.458 0.542)
 490) focus_area_special_needs>=0.5 106    36 completed (0.660 0.340) *
 491) focus_area_special_needs< 0.5 687   293 expired (0.426 0.574) *
 123) total_price_excluding_optional_support>=1.38e+03 3244  1210 expired (0.373 0.627) *
  31) focus_area_lit_language>=0.5 7492  2710 expired (0.362 0.638) *
```