Machine Learning Engineer Nanodegree

Capstone Proposal

Omar M. Adel Kamal

October 24th, 2019

Proposal

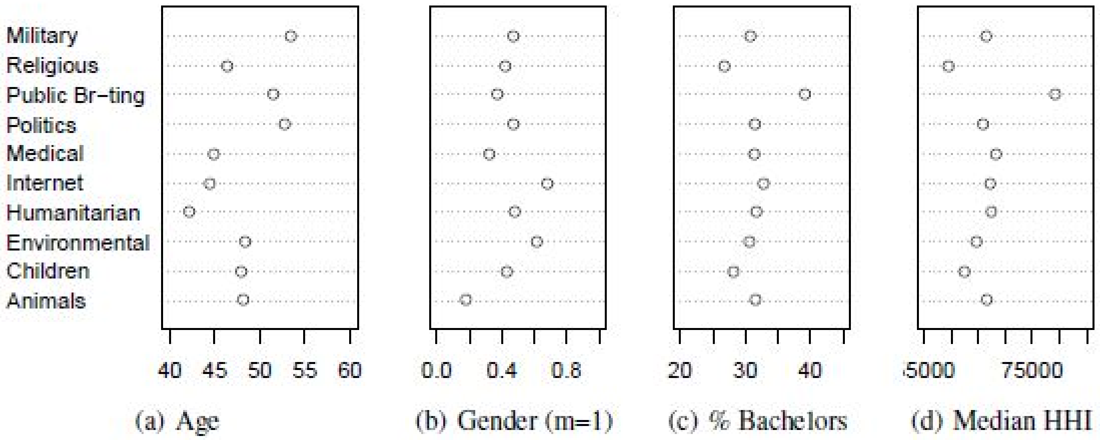
Domain Background

Founded in 2000 by a high school teacher in the Bronx, DonorsChoose.org empowers public school teachers from across the country to request much-needed materials and experiences for their students. At any given time, there are thousands of classroom requests that can be brought to life with a gift of any amount. DonorsChoose.org receives hundreds of thousands of project proposals each year for classroom projects in need of funding. Right now, a large number of volunteers is needed to manually screen each submission before it's approved to be posted on the DonorsChoose.org website.

Next year, DonorsChoose.org expects to receive close to 500,000 project proposals. As a result, there are three main problems they need to solve:

1. How to scale current manual processes and resources to screen 500,000 projects so that they can be posted as quickly and as efficiently as possible
2. How to increase the consistency of project vetting across different volunteers to improve the experience for teachers
3. How to focus volunteer time on the applications that need the most assistance

Related Academic Research
Some researches have been made to understand what makes a successful donation platform. They use game theory, machine learning and other knowledge domains. One of them is Giving is Caring:Understanding Donation Behavior through Email.



(a) Age (b) Gender (m=1) (c) % Bachelors (d) Median HHI

Figure 1. Demographics of donors to causes in 10 topical categories

Sociological studies on motivation often link motivation to group membership and identity definition. Demographics such as gender, age, and income, play an important role in an individual's formation of their self-image [39]. In order to determine if users' demographic characteristics play a role in their donation behavior, we collect the self-reported demographic information from the user profiles. This information includes age, gender, and zip code. In the case the user's zipcode is in US, we use census data to obtain localized statistics such as the percentage of bachelors degrees and median household income9. We aggregate these for the users donating to charities within the ten topics. The age, gender (male = 1), percentage of bachelor degrees and median household income are shown in Figure 1 (note, the 95% confidence intervals are too small to plot).

Several interesting trends become evident. First, note the average age is usually between 40 and 55 (even after removing the small number of instances where users claim age of > 100), showing that donors in our dataset tend to be of an older generation. Donors to political and military causes, as well as donors to public broadcasting, tend to be older. Environmental and internet (including sites like wikipeda.org) causes tend to attract slightly more male donors, whereas donors to animal-related causes tend to overwhelmingly be female. Public broadcasting, which is mostly operated by non-profits in US, attracts donors from neighborhoods with higher household income and a higher percentage of bachelordegrees.

My personal motivations are:

- I have always believed that the solution to my country's problems is education. Unfortunately, we don't have such a project or available data in my country but i believe if i can present a working example, someone from the Decision-makers in my country may respond positively.

- Our machine learning algorithm can help more teachers get funded more quickly, and with less cost to DonorsChoose.org, allowing them to channel even more funding directly to classrooms across the country.

Problem Statement

Our goal is to predict whether or not a DonorsChoose.org project proposal submitted by a teacher will be approved, using the text of project descriptions as well as additional metadata about the project, teacher, and school. DonorsChoose.org can then use this information to identify projects most likely to need further review before approval. To do so we will use either neural networks or ensemble of supervised classification models.

Datasets and Inputs

https://www.kaggle.com/c/donorschoose-application-screening/data

The dataset contains information from teachers' project applications to DonorsChoose.org including teacher attributes, school attributes, and the project proposals including application essays. Our objective is to predict whether or not a DonorsChoose.org project proposal submitted by a teacher will be approved.

Data Fields:

- id - unique id of the project application
- teacher_id - id of the teacher submitting the application
- teacher_prefix - title of the teacher's name (Ms., Mr., etc.)
- school_state - US state of the teacher's school
- project_submitted_datetime - application submission timestamp
- project_grade_category - school grade levels (PreK-2, 3-5, 6-8, and 9-12)
- project_subject_categories - category of the project (e.g., "Music & The Arts")
- project_subject_subcategories - sub-category of the project (e.g., "Visual Arts")
- project_title - title of the project

Note: Prior to May 17, 2016, the prompts for the essays were as follows:

- project_essay_1: "Introduce us to your classroom"
- project_essay_2: "Tell us more about your students"

- project_essay_3: "Describe how your students will use the materials you're requesting"
- project_essay_4: "Close by sharing why your project will make a difference"

Starting on May 17, 2016, the number of essays was reduced from 4 to 2, and the prompts for the first 2 essays were changed to the following:

- project_essay_1: "Describe your students: What makes your students special? Specific details about their background, your neighborhood, and your school are all helpful."
- project_essay_2: "About your project: How will these materials make a difference in your students' learning and improve their school lives?"

For all projects with `project_submitted_datetime` of 2016-05-17 and later, the values of project_essay_3 and project_essay_4 will be `NaN`.

- project_resource_summary - summary of the resources needed for the project
- teacher_number_of_previously_posted_projects - number of previously posted applications by the submitting teacher
- project_is_approved - whether DonorsChoose proposal was accepted (0="rejected", 1="accepted"); `train.csv` only

Proposals also include resources requested. Each project may include multiple requested resources. Each row in `resources.csv` corresponds to a resource, so multiple rows may tie to the same project by `id`.

- id - unique id of the project application; joins with `test.csv.` and `train.csv` on `id`
- description - description of the resource requested
- quantity - quantity of resource requested
- price - price of resource requested

Some features that may be interesting are school_state, as people will probably donate to schools in their state so states with more general income will probably have more donations and we want to accept proposals that will get more donations. Also project_subject_category as people will probably donate to categories they believe to be useful. Project_essays will be a critical factor as dedicated & serious applicants will always do their best to write a convenient essays to be accepted, so the more keywords

& length it contains the better it is. Project_resource_summary will be hard to process as there is a multitude of resources to think of, i think it will be easier if the resources were in terms of cash money.

Also we can notice that the data is imbalanced towards acceptance, the accepted projects are about 6 times larger than not accepted ones so we will have to do something about it, maybe we can drop a random portion of the accepted proposals.

Solution Statement

I will make a binary classification model that decides whether the proposal is approved or not. I will go through three phases: data discovery, data pre-processing & Feature engineering, model choosing then model implementing. I will try models like ensemble models, SVMs and other classification models based on the insights I gain from the data discovery phase.

Benchmark Model

Perhaps `teacher_number_of_previously_posted_projects` might provide a good signal as to whether a DonorsChoose application will be accepted? We can hypothesize that teachers who have submitted a large number of previous projects may be more familiar with the ins and outs of the application process and less likely to make errors that would lead to a rejection.

Let's test that theory by building a simple linear classification model that predicts the `project_is_approved` value solely from the `teacher_number_of_previously_posted_projects` feature. We will use logistic regression model from sklearn

Evaluation Metrics

Our goal is to predict whether an application to DonorsChoose is accepted. Submissions are evaluated on area under the ROC curve between the predicted probability and the observed target.

The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection in machine learning. The false-positive rate is also known as the fall-out or probability of false alarm and can be calculated as (1 − specificity).

Project Design

I intend to start with data discovery, i will read more about the domain and the meaning of the features. Then i will start with tokenizing strings, turning them all to lowercase, remove special symbols and i may need to extract numerical values in the resources as they may have an indication, also i might need to get words back to their origins to ease the process using stemming & lemming and make it more accurate then i will start extracting some features from the text features like searching for some key words if they exist in an essay or text matching using Pattern matching algorithms like Levenstien and Jaro-winkler or phonetic matching like Soundex and Metaphone, then i will start visualizing features and see which features are useful and which aren't. Then i will start shortlisting some models as NN based models and ensemble models, based on the insights i gain. Then i will train theses models using K-fold cross validation to see which model performs best based on the metrics used as recall & precision, then i will tune the hyper-parameters of that model using gridsearchcv.

References

1-https://www.kaggle.com/c/donorschoose-application-screening/overview
2- https://www.donorschoose.org/about
3- https://en.wikipedia.org/wiki/Receiver_operating_characteristic