ISyE 6740 - Fall 2021
Project Proposal

Team Member Name: Chai Han Xie (hchai30)

Project Title: Project Recommender for a Donation Website

Problem Statement

DonorsChoose.org is an online website that helps to link donors with projects to help students in need. These projects are submitted by teachers, alongside information such as the need statement and the materials needed. When a project reaches its funding goal, DonorsChoose.org will ship the requested materials to the school.

Over the years, the number of projects submitted to the website have been rising steady (see Figure 1), making it increasingly difficult for donors to sift through all the projects to find something that they want to support. Even with a filter function, donors might have to look through a long list for popular subjects or areas, and hence could be discouraged from donating.

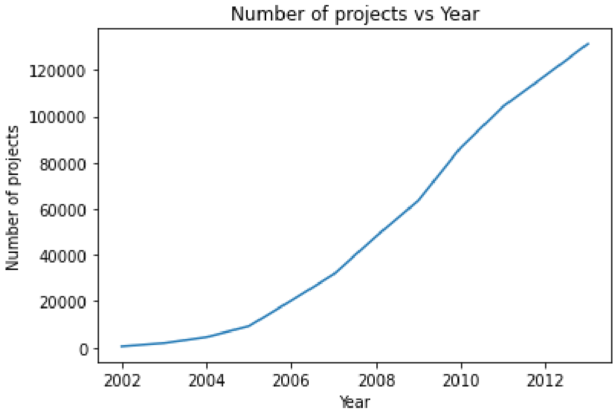

Figure 1: Number of projects on DonorsChoose.org have been increasing over the years.

It is therefore imperative to develop a recommender to identify projects that are of exceptional quality and display these projects to donors on the website. This would hopefully improve the number of donations, the number of successfully funded projects, and most importantly the number of students that will benefit from the donations.

To help with this, DonorsChoose.org has identified a subset of projects that they term as *exciting*. These projects has meet the following requirements.

- The project was fully funded.

- At least one donor donated to the project because of a shared link by the teacher. This type of donors will be referred to as teacher-referred donors.
- The comment thread of the project had a higher than average percentage of donors leaving an original message.
- The project received a donation via credit card, PayPal, Amazon, or check.
- In addition to every requirement above, the project also met one or more of the following requirements:
  - The project received donations from three or more teacher-referred donors.
  - A non-teacher-referred donor donated more than $100.
  - The project received a donation from a list of around 15 donors. The list was curated by DonorsChoose.org.

Data Source

There are three main dataset, each containing different information about a project.

essays.csv contains the essays and text provided by the teacher when submitting a project onto DonorsChoose.org. The data fields in **essays.csv** can be found in Table 1.

projects.csv contains most of the information about the project (e.g. resource type requested, total price of project), the teacher (e.g. prefix, qualifications), and the school that the teacher is from (e.g. city school is in, poverty level of school). The data fields in **projects.csv** can be found in Table 2.

outcomes.csv contains the requirements that DonorsChoose.org specified for exciting projects. This includes a *is_exciting* feature that aggregates the requirements. The data fields in **outcomes.csv** can be found in Table 3.

The three dataset can be obtained from Kaggle (*https://www.kaggle.com/c/kdd-cup-2014-predicting-excitement-at-donors-choose/data*).

Methodology

We will first focus on the **projects.csv** dataset as input data, given that the data is structured and there is a notable number (24) of features available.

In addition, no projects submitted before 2010 is labelled by DonorsChoose.org as exciting . As such, we shall omit all projects before 2010, which still leaves us with a sizable amount of 484,371 projects (from 664,098 projects).

The remaining sections in this report will be as follows:

- **Feature selection and engineering**: We will identify important features that could be useful in predicting if a project is exciting, and reduce the number of features to a more manageable size. We will also explore whether it is possible to obtain more useful features by combining features, extracting information from existing features (e.g. month from the date posted(, or creating new features (e.g. the number of past projects from the same teacher that have been labelled as exciting).
- **Model fitting and evaluation**: After we have identified useful features, we will use these features as inputs to classification models that we have learnt in this course, such as logistic regression, support vector machine, and neural networks. We will then compare their accuracy for predicting whether a project is exciting or not.

If time permits, we will assess two potential improvements.

- We can explore if features extracted from the unstructured data in **essays.csv** can help to improve the model.

- Rather than fitting a single model to predict *is_exciting*, we can create multiple sub-models to predict the requirements for exciting projects instead. The output from each model could then be combined to form an overall prediction of whether the project is exciting.

Feature Selection and Engineering

To have a sense of the original features' usefulness, the mutual information (MI) of each feature with the label *is_exciting* is computed and shown in Figure 1.

MI of each feature X_i with the label Y quantifies the amount of information obtained about Y or the reduction in uncertainty of Y by observing X_i. Unlike measures such as correlation coefficient, MI is not limited to linear dependence, and can be computed from the entropy of Y ($H(Y)$) and the conditional entropy of Y given X_i ($H(Y|X_i)$).

$$I(X_i, Y) = H(Y) - H(Y|X_i)$$

Figure 2 shows that most of the original variables provide little information about the label. Among the variables, the cost of fulfilment, which measures the project's price excluding any portion of the donation that goes to supporting DonorsChoose.org, provides the highest amount of MI. This is followed by features about the school's location, such as latitude/longitude, district, city, county, and state.

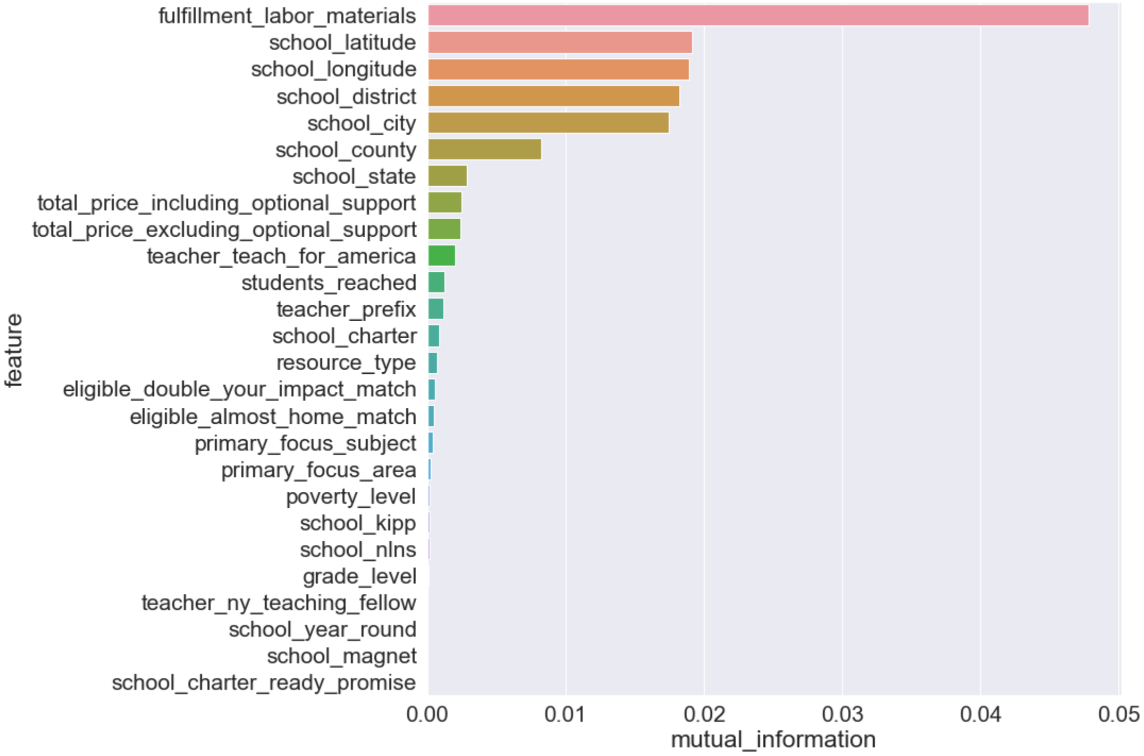

Figure 2: Mutual information against label *is_exciting*.

Given that features about the project, teacher, and schools seem to provide little information, the next step will be to explore whether it is possible to obtain more useful features by creating new features.

Model Fitting and Evaluation

For model evaluation, we will split the data into the training set (for model training) and the test set (for measuring performance). Instead of a random split, we will use the latest 20% as the test set, to assess both over-fitting and possible time drift.

The following evaluation metrics will be used to evaluate the models:

- **Precision**: A model with a higher precision means that there is a lower chance that the model will recommend a project that is not exciting.

$$\text{Precision} = \frac{\text{Number of exciting projects that are correctly identified by the model}}{\text{Total number of projects that the model identified as exciting}}$$

- **Recall**: A model with a higher recall means that the model has a lower chance of missing out on exciting projects.

$$\text{Recall} = \frac{\text{Number of exciting projects that are correctly identified by the model}}{\text{Total number of exciting projects}}$$

- **F1-score**: The F1-score can be interpreted as a weighted average of the precision and recall. Similar to precision and recall, a better performing model has a higher F1-score.

$$\text{F1 score} = \frac{2(\text{Precision})(\text{Recall})}{\text{Precision} + \text{Recall}}$$

- **Area Under the Curve of the Receiver Characteristic Operator (AUC-ROC)**: The AUC-ROC captures the chance that the model gives higher probability to a randomly selected exciting project compared to a randomly selected non-exciting project. The higher the AUC-ROC, the better the performance of the model at distinguishing between exciting and non-exciting projects.

Data field	**Description**
projectid	Unique id for each project
teacher_acctid	Unique id for the teacher who created the project
title	Text field for the title of the project
short_description	Text field for a short description of the project
need_statement	Text field for the need statement of the project
essay	Text field for an essay about the project

Table 1: Data fields in **essays.csv**.

Data field	**Description**
projectid	Unique id for each project
teacher_acctid	Unique id for the teacher who created the project
schoolid	Unique id for the school where the teacher is from
school_latitude	Latitude of school
school_longitude	Longitude of school
school_city	City of school
school_state	State of school
school_zip	Zip of school
school_metro	Metro of school
school_district	District of school
school_county	County of school
school_charter	Whether school is a public charter school
school_magnet	Whether school is a public magnet school
school_year_round	Whether school is a public year round school
school_nlns	Whether school is a public nlns school
school_kipp	Whether school is a public kipp school
school_charter_ready_promise	Whether school is a public ready promise school
teacher_prefix	Prefix of teacher
teacher_teach_for_america	Whether teacher is from Teach for America
teacher_ny_teaching_fellow	Whether teacher is from New York City Teaching Fellows
primary_focus_subject	Main subject for the project
primary_focus_area	Main area for the project
secondary_focus_subject	Secondary subject for the project
secondary_focus_area	Secondary area for the project
resource_type	Main type of resource requested by the project
poverty_level	Poverty level of school
grade_level	Grade level for which project materials are intended
fulfillment_labor_materials	Cost of fulfilment
total_price_excluding_optional_support	Project cost excluding optional tips by donors
total_price_including_optional_support	Project cost including optional tips by donors
students_reached	Number of students benefiting from the project
eligible_double_your_impact_match	Whether a corporate partner is supporting half of the project cost
eligible_almost_home_match	Whether a corporate partner is sponsoring US$100
date_posted	Date the project was submitted

Table 2: Data fields in **projects.csv**.

Data field	**Description**
is_exciting	Label provided by DonorsChoose.org if the project is exciting
at_least_1_teacher_referred_donor	Boolean variable whether a donor donated because of a shared link by the teacher
fully_funded	Boolean variable whether the project was successfully funded
at_least_1_green_donation	Boolean variable whether a donation is made with credit card, PayPal, Amazon, or check
great_chat	Boolean variable whether the comment thread for the project has greater than average unique comments
three_or_more_non_teacher_referred_donors	Boolean variable whether there are at least three donors that landed on the site by means other than a teacher referral link/page
one_non_teacher_referred_donor_giving_100_plus	Boolean variable whether a donor that landed on the site by means other than a teacher referral link/page donated more than US$100
donation_from_thoughtful_donor	Boolean variable whether a project received a donation from a curated list of around 15 donors from DonorsChoose.org

Table 3: Data fields in **outcomes.csv**.