Machine Learning Engineer Nanodegree
Capstone Proposal

Bharadwaz Mahankali

March 06, 2018

Domain Background

DonorsChoose.org receives hundreds of thousands of project proposals each year for classroom projects in need of funding. 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 that need to be solved:

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

Problem Statement

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.

Datasets and Inputs

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

File descriptions

train.csv : The training set.
test.csv : The test set.
resources.csv : Resources requested by each proposal; joins with test.csv and train.csv on id.
sample_submission.csv : A sample submission file in the correct format
Data fields.

test.csv and train.csv :

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.
project_essay_1 - First essay.
project_essay_2 - Second essay.
project_essay_3 - Third essay.
project_essay_4 - Fourth essay.
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.

resources.csv :

Proposal 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 test.csv and train.csv on id.
description - Description of the resource requested.
quantity - Quantity of resource requested.
price - Price of resource requested.

A note on essay data

Prior to February 18th, 2010, for their DonorsChoose application, teachers had the option of writing either a free-form essay (split into project_essay_1, project_essay_2, project_essay_3, and project_essay_4) or writing free-form answers to the following four prompts:

Introduce your classroom (project_essay_1)
Describe the situation (project_essay_2)
Describe the solution (project_essay_3)
Empower your donors (project_essay_4)

Solution Statement

Machine Learning Ensemble methods that perform best will be used for classification based on which decision whether proposal has to be approved can be made. Since essays written by teachers might play significant role in decision making, NLP techniques(provided by sklearn) will be used to effectively analyze essays and other text data. Goal is to bring Area under the ROC curve score, which is used for evaluation, closer to 1.

Benchmark Model

Kaggle DonorsChoose.org Application Screening competition leaderboard score 0.79 (Area under receiver operating characteristic curve) will be used as benchmark score. An attempt will be made to stand on top 30% of public leaderboard.

Evaluation Metrics

Model will be evaluated on area under the ROC curve between the predicted probability and the observed target.

Project Design

Data Exploration : Investigating DonorsChoose.org to better understand the relationship between different data elements.

Data Preprocessing : Clean and format data that can be used as input for ML algorithms. Transform text data using NLP techniques.

Model Selection : Use various ensemble algorithms and choose the model that performs best.

Model Tuning : Tune parameters of model for better performance.

Model Evaluation : Test the model by submitting the output file as specified in Kaggle competition.

Conclusion : Provide the results after model evaluation and discuss learnings.