Machine Learning Engineer Nanodegree

Capstone Proposal

Anirudh Varanasi
June 05th, 2018

Title

DonorChoose.org Application Screening

Proposal

Domain Background

DonorChoose.org is a non-profit organization, based in the Bronx was founded by a former public school teacher Charles Best. This organization empowers public school teachers from across the country to request much-needed materials and experiences for their students. Teachers submit project proposals to DonorChoose.org through its website. DonorChoose.org then manually screens and approves the proposals before posting on the website. Right now, a large number of volunteers are needed to manually screen each submission before it is approved by DonorChoose.org. Next year, DonorChoose.org expects to receive close to 500,000 project proposals. The goal is to predict whether or not a DonorChoose.org project proposal submitted by a teacher will be approved.

One popular classification example is sentiment analysis where class labels represent the emotional tone of the source text such as “positive” or “negative”. Other examples widely seen are

- Spam filtering – classifying email text as span or not[1].
- Sentiment analysis of rotten tomato movie reviews[2]
- Sentiment analysis of amazon product reviews. IMDB movie reviews and topic categorization of news articles[3].
- Sentiment analysis of movie reviews, classifying sentences as being subjective or objective, classifying question types, sentiment of product reviews and more[4].
- Readability assessment, automatic determining the degree of readability of text, either to find suitable materials for different age groups or reader types or as part of a larger text simplification system.
- Sentiment analysis, determining the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document.
- Health-related classification using social media in public health surveillance[6].
- Article triage, selecting articles that are relevant for manual literature curation, for example as is being done as the first step to generate manually curated annotation databases in biology[7].

Problem Statement

Next year, DonorChoose.org expects to receive close to 500,000 project proposals. As a result, manual screening of application would be a tedious task. Using machine learning, through this project an attempt to discover a classification model that predicts which application to accept or reject. There are three main problems they need to attention:

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.

Building a model to predict a proposal's approval accomplishes all three of these goals in a quantifiable, measurable and replicable way.

Datasets and Inputs

The dataset can be downloaded from kaggle.com. The dataset contains information from teacher's project applications to DonorsChoose.org including teacher attributes, school attributes, and the project proposals including application essays.

- train.csv - The training set contains the label indicating whether the proposal was approved
- test.csv - The test set contains information without label indicating the proposal was approved.
- resources.csv – The resources set contains description, quantity and price of resources requested as part of the dataset.
- 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

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.

The training dataset could be split into 80 train and 20 validation sets.

The data fields that might be useful in predicting the project_is_approved (label) are:

- teacher_prefix,
- school_state,
- project_submitted_datetime,
- project_subject_categories,
- project_grade_categories,
- project_subject_subcategories,
- project_title,
- teacher_number_of_previously_posted_project

resources.csv:

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

Solution Statement

The most common solution to such problems is the method of classification. Some of the classification methods are:

a. Logistic Regression
b. Random Forest
c. Naïve Bayes
d. Support Vector Machines

Another alternative is to use Deep learning with Tensorflow/Keras.

Benchmark Model

At the present DonorChoose.org manually screens for applications and also the model mentioned in the tutorial “Getting Started with the DonorsChoose.com Data set” by Sanders Kleinfeld uses linear classification model to train. The training and validation log losses are the results which are the outputs. Later, AUC (area under the curve), which is the metric for assessing the accuracy of prediction calculated. The model achieves AUC score of 0.56.

A classification model, which will screen for applications quickly than the manual process would solve the problems of DonorChoose.org

Evaluation Metrics

Metrics used to evaluate is AOC (area under the ROC curve between the predicted probability and the observed target). 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 knows as sensitivity, recall or probability of detection. The false positive rate is also known as the fall-out or probability of false alarm.

Project Design

From the description and problem statement, it can be concluded that a classification algorithms like Naïve Bayes, Random Forest and Neural Networks can be used to arrive at a solution. The steps for this are as follows

- Data Exploration, Visualization, Preprocessing, Feature Engineering:
    - Visualizing the dataset to find the degree of correlations between predictors and target variable and the related predictors are picked from the observations,
    - Detecting outliers,
    - Removing null values,
    - Cleaning the dataset,
    - Checking relevant columns,
    - Engineering features: Information from EDA, could provide use new insights if new features could be created.
        - Example: The month from 'project_submitted_datetime' can be used to compare and see if there is any correlation.
        - Example: From the resources dataset, the price, number and can be checked to see if there is any correlation.
        - Example: The time at which the proposal was submitted can also be important. Likewise, day of the week, for this decision trees would be helpful.
- Model Selection and Model Tuning: Considering multiple supervised ML models like Logistic regression, Naïve Bayes, SVM, Random Forest, Gradient Boosting and Neural Networks.
- Testing and Optimizing: Optimizing the model

References:

1. http://airccse.org/journal/jcsit/0211ijcsit12.pdf
2. https://www.cs.umd.edu/~miyyer/pubs/2015_acl_dan.pdf
3. https://arxiv.org/abs/1412.1058
4. https://arxiv.org/abs/1408.5882
5. https://ieeexplore.ieee.org/document/7925400/
6. https://ieeexplore.ieee.org/document/7925400/
7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2559988/