Capstone Proposal
Udacity Machine Learning Engineer Nanodegree

Aaron Penne
2018-05-19

1 Domain Background

Teachers in low income schools often find the school is unable to provide their students with adequate resources. This leads some teachers to purchase items with their own money to meet the students' needs. To help solve this issue, Charles Best created the website DonorsChoose.org to allow philanthropic individuals the opportunity to selectively donate to various teacher projects. Teachers submit their applications to DonorsChoose.org, and if they are accepted then the project goes public and people may donate to it.

This charity has served as a sort of Kickstarter for small public school needs and provides a direct way to make an impact on young people's lives. DonorsChoose.org has received top industry awards every year since 2005 and has fulfilled over 600,000 classroom projects. This impact can be made even greater by leveraging advanced technologies to actively utilize the datasets on hand.

2 Problem Statement

DonorsChoose.org must read through hundreds of applications in detail, employing many volunteers to do so. This takes a lot of time and resources for the organization, and also adds delay to the teachers waiting for approval. By automating a large segment of this process, applications that have an obvious classification can be accepted/rejected, while the more nuanced applications can be read more in depth by a volunteer. The problem to be solved is a classification one, should the DonorsChoose.org accept or reject a given proposal.

3 Datasets and Input

The dataset is compiled from historical project applications to DonorsChoose.org. Each request contains metadata such as the school's state, class grade, teacher ID, categories, submission datetime, etc. The primary content of each request is essays filled out by the teacher, with 2-4 essays per request. Another resources dataset is provided that details the description, quantity, and price of each item in the applications. This resources dataset is tied back to the applications dataset using a unique id for each application.

The training set has 182k records and the testing set has 78k records. There are two class labels in this dataset, "accepted" and "rejected". An interesting break in the data is that proposals before a certain date had 4 questions to be answered. The later proposals had 2 questions. It will take some experimentation to determine how to properly handle this discrepancy, more information on this is in Section 7

Many of the attributes are numerical in nature. These will be the simplest to model. Many attributes are also categorical, meaning they can be converted to numerical values and modelled in similar ways. The bulk of the potentially interesting data, however, is the essays written by the teachers. These provide both a challenge and an opportunity, as there is quite a bit of information encoded in these texts.

4 Solution Statement

This automated system will incorporate two models. One model will use some neural network implementation with the categorical/numerical values as inputs. Simple mapping of the data to

one-hot and running through a custom neural network will hopefully yield useful results before incorporating the essay information.

Another model will be based on natural language processing (NLP) techniques such as topic modelling to compare the accepted essays to the rejected essays. There are multiple techniques available to do this, with doc2vec being a primary contender for this application. These two approaches will be incorporated into an ensemble method, with more weight given to the better performing of the two models.

5 Benchmark Model

Random selection will be the first model to compare against, weighted with the same class distribution as the training set. The training set is imbalanced and has a significantly larger number of accepted proposals, so that percentage will be used for the random selection. This model is to be a naïve baseline to get a sense of where zero performance lies.

K nearest neighbors (KNN) will also be used to attempt classification with the minimum number of informative variables. If the KNN model performs reasonably well, then the neural network implementation model should perform even better. In particular, the ensemble method combining neural network and NLP implementations should greatly outperform the KNN implementation.

6 Evaluation Metrics

For two class classification problems, a confusion matrix contains actual and predicted results. A confusion matrix has four values: True positive (TP), true negative (TN), false positive (FP), and false negative (FN). These metrics use “positive” to denote an accepted proposal, and “negative to denote a rejected proposal. A true positive refers to the count of proposals that were accurately predicted to be accepted. A false negative refers to the count of proposals that were incorrectly predicted to be rejected.

Several metrics are derived from the confusion matrix, and these will ultimately be the comparison metrics between models. They are sensitivity, specificity, accuracy, geometric mean, and F score. It is likely that the F score will be the single metric used most to measure performance as it incorporates precision and recall. These measure exactness and completeness respectively, by calculating ratios between TP, FP, and FN.

A receiver operating characteristic (ROC) curve can be used to visualize some of the information displayed in a confusion matrix. The ROC curve is computed by varying the threshold between TP and FP, and it visually conveys the ability of the system to differentiate between the two. The area under the ROC (AUC) can be computed as a summary statistic. The AUC will be used in conjunction with the F score.

7 Project Design

This project will be completed with Python 3.6. Libraries involved will be scikit-learn, pandas, matplotlib, NLTK, Keras, and Tensorflow, and possibly more depending on need. The outputs of the neural network and topic modelling models will be part of the final ensemble method. Depending on the KNN performance, that may be included as well.

Before beginning investigations, the data will be pre-processed. The resource data will need to be mapped to the proposal data. An interesting problem to deal with in this dataset is that before a certain date the proposals included 4 questions. After that date the newer proposals included 2 questions. This means that something needs to be done with the NaN values present in the newer data. There are a few options, and each will be tested to see if a positive or negative impact on the score occurred. The common thing to do when an attribute has a large percentage of missing data is to simply delete that attribute. This seems to be an erroneous approach in this case, as much data will be lost. Reading the proposal prompt questions shows that the old prompts #1 and #2 are very similar to the new prompt #1. Also, the old prompts for #3 and #4 are very similar to the new prompt #2. This leads to another option of melting the old #1 and #2 questions into #1, and old #3 and #4 into #2. Alternatively, all of the text can be combined into a single text field then processed.

Initial investigations will include the rate of acceptance of different attributes (% accepted per state for example) and correlation of various attributes. This stage will also involve visualizing the data to determine which variables provide information and which ones only contribute noise.

The first pass of this project will be an attempt to classify the proposals based solely on metadata and resources requested. The essays will be ignored. This will allow for a quick prototype of a classifier without diving into the complexities of text. The second pass will only use the essays as the primary inputs. NLP techniques such as topic modelling will be used to compare the essays of accepted and rejected proposals. Finally, an ensemble method combining these models will be used to make the final prediction.

8 References

M. Fatourechi, R. K. Ward, S. G. Mason, J. Huggins, A. Schlögl and G. E. Birch, "Comparison of Evaluation Metrics in Classification Applications with Imbalanced Datasets," *2008 Seventh International Conference on Machine Learning and Applications*, San Diego, CA, 2008, pp. 777-782.