Predictive Analytics of Donors in crowdfunding platforms: A case study on Donorschoose.org

Joel Varma Dirisam
Department of Computer Science
California State University, Fullerton
Fullerton, USA
joelvarma@csu.fullerton.edu

Doina Bein
Department of Computer Science
California State University, Fullerton
Fullerton, USA
dbein@fullerton.edu

Abhishek Verma
Department Computer Science
New Jersey City University
New Jersey, USA
av56@njit.edu

***Abstract*— Every year donors support crowdfunding organizations like DonorsChoose.org. With the decline in the number of contributions received every year, initiatives in United States schools are affected. Donorschoose is a readily available dataset which contains data about US school projects proposed by teachers. In this paper, we apply machine learning algorithms, and we analyze each feature of the Donorschoose including the actions of donors for every project to encourage school's funding. We also provide probability and statistical modeling techniques to compare with the machine learning results. Some of the findings include that donor will be retained if they had a positive interaction overall with the organization, and the targeting donors specific to location also improves new donors to join. Finally, we present some main facets of the dataset concerning the approval of the project donations.**

Keywords—Crowdfunding, machine learning algorithm, natural language processing, NLP probabilistic model, statistical model.

I. INTRODUCTION

Every year donors support crowdfunding organizations like Donorschoose.org. With the decline in the number of contributions received every year, initiatives in United States schools are affected. Crowd-sourced donations or crowdfunding offers an innovative approach of fundraiser. Online crowdfunding sites including Kickstarter.com or DonorsChoose.org allow people to publish initiative requests and collect funds from public for innovative products, promote creative and research projects, and most importantly public-school education. Organizations like these made a direct impact on school learning from 2005. In 2012 alone, this organization has a funding of $2.7 Billion.

However, engaging the donors and retaining them is the key to the success of this idea. There has been a donor attrition and only 20-25% donors are interested to make more than one donation. Maintaining a proper relation with existing donors is obviously effective approach to retain the donors, this is very crucial because a slight improvement in the donor retention can mean the donations could yield almost significant increase in the donations. There are still some facts regarding donors' behavior which are not very well understood.

DonorsChoose.org has funded over 1.1 million classroom requests through the support of 3 million donors, the majority of whom were making their first-ever donation to a public school. If DonorsChoose.org can motivate even a fraction of those donors to make another donation, that could have a huge impact on the number of classroom requests fulfilled.

Donorschoose is a readily available dataset [11, 12], which contains data about US school projects proposed by teachers with the goal of making the classroom learning better for students. The dataset also contains users teachers and donors' data, which helps in prediction of the behavior to improve the retention rates.

In this paper, we apply machine learning algorithms, and we analyze each feature of the Donorschoose dataset [11, 12], including the actions of donors for every project to encourage school's funding. We also provide probability and statistical modeling techniques to compare with the machine learning results. Some of the findings include that donor will be retained if they had a positive interaction overall with the organization, and the targeting donors specific to location also improves new donors to join. Finally, we present some main facets of the dataset concerning the approval of the project donations.

The proposed system will address the problem of retaining the existing donors by applying the machine learning models to analyze the behavior of the donors and filter the most promising projects from the dataset. The primary objective to apply machine learning is to extract 10 most important factors for a project to be picked by a donor. This will give a better idea to school teachers so that they can formulate their projects in more winnable ways. These features can also be crucial to the donors to keep them donating to the school teachers.

The paper is organized as follows. In Section II we present the background and related work, followed by the system architecture in Section III. Simulations and results are presented in Section IV. Concluding remarks and future work are given in Section IV.

II. BACKGROUND AND RELATED WORK

The authors in [1] noted that communicating with the donors has maintained the retention rates. The authors of [2] concluded that a notable aspect of crowdfunding is the large regional distribution of donors into tiny, early-stage ventures. This conflicts with current models that would forecast the co-location of entrepreneurs and investors due to distance-sensitive costs [2].

The authors in [3] have concluded that efforts to develop technology to support the formation of viable and effective online communities can do more than simply provide access to an infrastructure that allows ongoing group discussion to be shared and structured.

The DonorsChoose.org accepts more than a million project proposals every year. Each of these projects needs to be picked by donors considering for a donation. This huge number of project proposals are overwhelming for the donors and needs a way to find subset of significant projects. The project essays describe the type of project and how it helps students. These essays contain essential information which primarily helps in project acceptance.

A good solution will enable DonorsChoose.org to build targeted email campaigns recommending specific classroom requests to prior donors. Part of the challenge is to assess the needs of the organization, uncover insights from the data available, and build the right solution for this problem. Submissions will be evaluated on the following criteria:

- Performance - How well does the solution match donors to project requests to which they would be motivated to donate? DonorsChoose.org will not be able to live test every submission, so a strong entry will clearly articulate why it will be effective at motivating repeat donations.

- Adaptable - The DonorsChoose.org team wants to put the winning submissions to work, quickly. Therefore, a good entry will be easy to implement in production.

- Intelligible - A good entry should be easily understood by the DonorsChoose.org team should it need to be updated in the future to accommodate a changing marketplace.

Machine learning is about building computer systems or programs that can learn from data.

Non-neural network approaches such as Support Vector Machines (SVM) and Naive Bayes have also been used with some success for sentiment analysis. While being versatile, they are not as powerful as neural networks, especially deep learning networks. Deep learning networks are powerful machine learning algorithms that require massive amount of data to achieve high performance or accuracy. Once an accurate model is developed, users will be able to feed sequential data of fixed length to the model and be able to get the output as the prediction for the next ten days.

We present briefly some of the machine learning algorithms we will use to compare each machine learning model performance.

K-Nearest Neighbor (K-NN) is a distance-based algorithm which is widely used for problems which needs classification/regression. The 'K' is a hyperparameter is generally an odd number to let the machine pick which particular class has most data. This is called a model centered on distance, as it uses Euclidean distance as a metric to classify two data points. Another metric used to address textual data is hamming distance.

Naïve Bayes classifier [9] is based on Conditional Probability theorem by Bayes scientist hence implying the attributes are all independent. Naïve Bayes can yield higher accuracies for most of the problems because the naïve assumption it makes while computing the conditional probability. Refer the below conditional probability formula to know how each feature probability against the class is computed:

$$\begin{aligned} p(C_k, x_1, \dots, x_n) &= p(x_1, \dots, x_n, C_k) \\ &= p(x_1 \mid x_2, \dots, x_n, C_k)\, p(x_2, \dots, x_n, C_k) \\ &= p(x_1 \mid x_2, \dots, x_n, C_k)\, p(x_2 \mid x_3, \dots, x_n, C_k)\, p(x_3, \dots, x_n, C_k) \\ &= \dots \\ &= p(x_1 \mid x_2, \dots, x_n, C_k)\, p(x_2 \mid x_3, \dots, x_n, C_k) \dots p(x_{n-1} \mid x_n, C_k)\, p(x_n \mid C_k)\, p(C_k) \end{aligned}$$

Logistic Regression (LR) model is probabilistic-based and uses a logistic function to model some binary variable but it could have a lot more target labels [5]. So one uses one versus other modeling method. LR finds a high dimensional space hyper plane by accurately distinguishing the classes. LR is more about finding a hyperplane in high dimensional space which can linearly separate the data points according to their classes. This hyperplane can be represented by a matrix, WT. This WT needs to be multiplied by X which is an arbitrary data point to check the class of X. Cosine multiplication of WT * X results in a value greater than 0 if X is above hyperplane and a value less than 0 if X is below hyperplane.

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that work by building a multitude of decision trees at training time and producing classes that are class mode (classification) or mean prediction (regression) of individual trees. Random decision forests are right for decision trees' overfitting habits. Tin Kam Ho created the first algorithm for random decision forests using the random subspace method which, in Ho's formulation, is a way of implementing Eugene Kleinberg's proposed "stochastic discrimination" approach to classification.

Neural networks are the usual representation we make of the brain based on nature: neurons that are interconnected to other neurons that form a network. Before being an actual item, a basic detail transits in many of them, like "moving the hand to pick up the pencil.

Donorschoose is a readily available dataset which contains data about US school projects proposed by teachers. It contain columns that are of different size or range. Some features may have a range of 0-10 and another may have a range of 2000-100000. Because of the large-scale difference in size, normalization and standardization of the data are important steps in improving the quality of the data thereby effecting the accuracy.

Sentiment Polarity is a sentiment value for a given English sentence; the score with 1 is very positive while 0 is very negative. We used this technique also to see if the model trains well.

Global Vectors is a library that is freely accessible, created by researchers from Stanford. For each word, this library contains 300 dimensional vectors. It is a cosine angle based mathematical vectors for every word. Bag of Words is a basic method for converting data from a sentence to an n-dimensional vector. However, this model does not care about the order of appearance of the words hence loses the meaning.

CountVectorizer(CV) is used change the columns of selected data into vectors Bag Of Words model, TFIDF. Applying these vectorizations can be increasing the number of columns exponentially but we can lessen it by using Principal component Analysis (PCA). PCA is a dimensionality reduction technique to reduce the data features after pre-processing.

Term frequency and Inverse Document Frequency (TFIDF) is based on the frequency of terms against overall data and dividing it with log based inverse document frequency to weigh down unimportant words. This helps in rating the importance of the word corresponding to that document.

ROC-AUC value is considered when accuracy might be flawed because of the unbalanced dataset. Typically, AUC value starts from 0.5 and tends to 1.0 (optimal).

III. System Architecture

The project is completely developed using Jupyter which is a part of Anaconda distribution. Anaconda is an open-source distribution containing all the Python packages required. The project consists of some preprocessing modules which help in cleaning the text and applying other natural language processing techniques and also for vectorization of words with Glove model. The system architecture and the flow of processes is shown in Fig. 1. Since we need to apply different machine learning algorithms, converting these routine tasks into separate modules helps in reusing the code.

The input features for the machine learning algorithms are: Project title, Project short essay, Project categories, Project subcategories, Project submitted Date Time, Project grade category, number of projects proposed by teacher, Donor name, Donor State, Donation time details, School name, School metro type, School state, School Zip, School county, School city, Teacher prefix, Teacher first project proposed time, Resource items, Resource quantity, Resource unit price, Resource vendor name, Donor details etc.

The output is binary since we consider only a binary classification with two classes: Project is approved or Project rejected. Two of the essential strategies for data cleaning are the date and time extraction. While this value column (2016-12-05 13:43:57) may seem not useful, sometimes it the machine can use only the year and month instead of all the minute details. Extracting year and month helped the model accuracy improve by 1% which is very significant for such a small step of preprocessing. It is also important to bin the number based features because it only makes machine learning model harder to predict continuous range values of numerical values. It is a very basic pre-processing step.

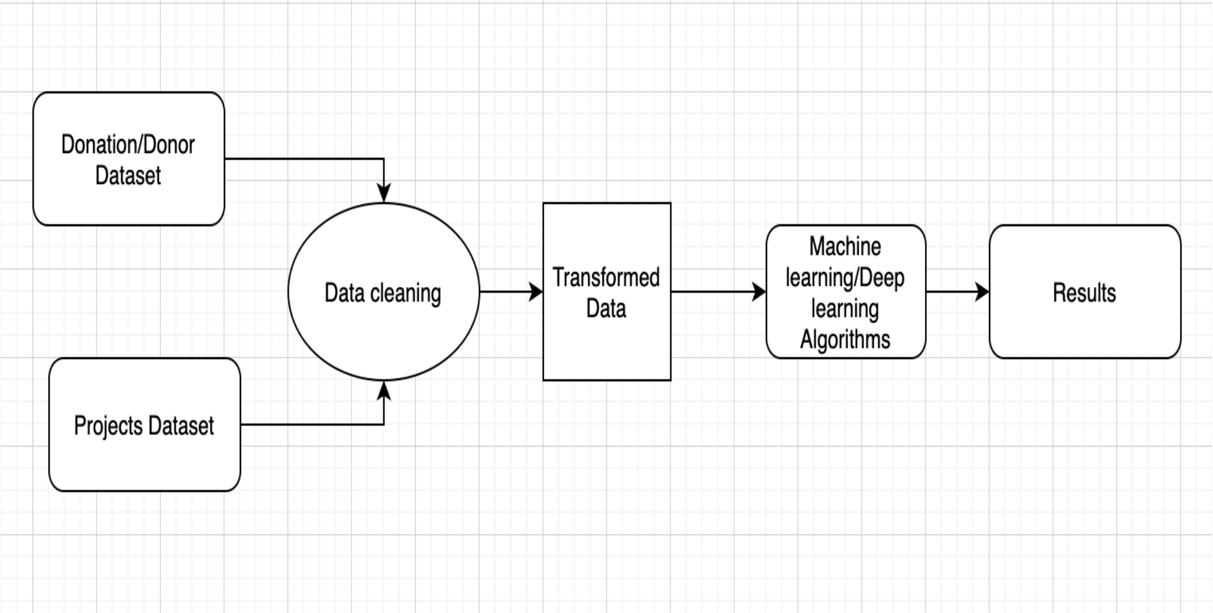


Figure 1. System Architecture/ Process Flow

IV. Simulations and Results

The outcomes of KNN simulation results are shown in Figures 2 and 3. KNN has slightly increased the Area Under Curve value and it is not useful for BOW or TF-IDF vectorization because the memory limitations restrict the size of the dataset. K-NN is useful where scale or dimensions of the dataset are smaller (see Fig. 4).

Naïve Bayes performed much better with Bag of words vectorizer (see Fig. 5) than with TF-IDF (see Fig. 6) due to the fact that BOW is simple and the conditional probability computations have not been overfitted, but it is otherwise for the TFIDF (see Fig. 7).

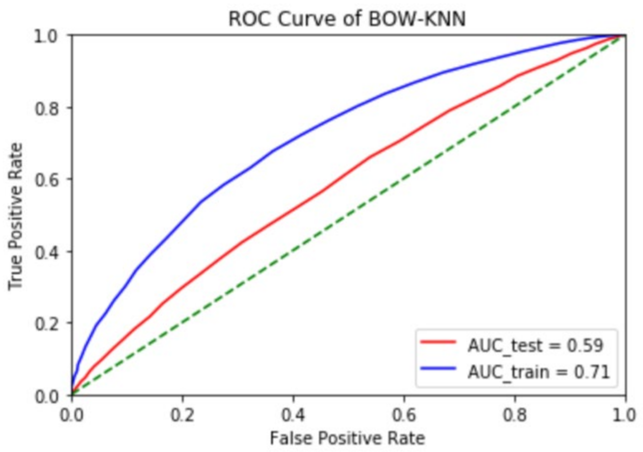


Figure 2. AUC Curve of Bag of Word-KNN

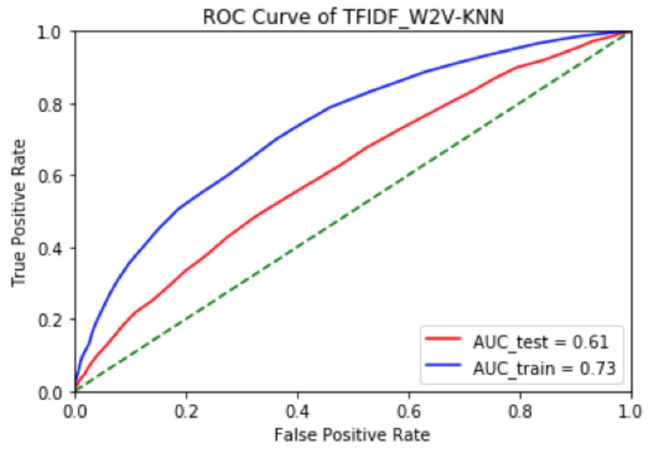
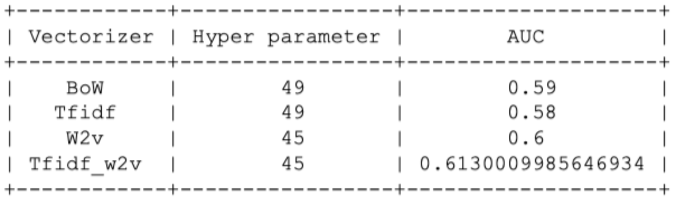


Figure 3. AUC Curve of TFIDF-KNN



```
+------------+-----------------+----------------------+
| Vectorizer | Hyper parameter |         AUC          |
+------------+-----------------+----------------------+
|    BoW     |       49        |         0.59         |
|   Tfidf    |       49        |         0.58         |
|    W2v     |       45        |         0.6          |
| Tfidf_w2v  |       45        |  0.6130009985646934  |
+------------+-----------------+----------------------+
```

Figure 4. Comparison of different vectorizers

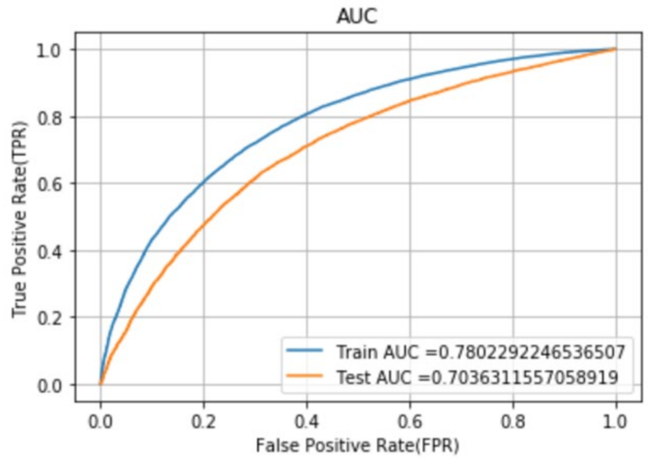


Figure 5. AUC of Bag Of Word- Naïve Bayes

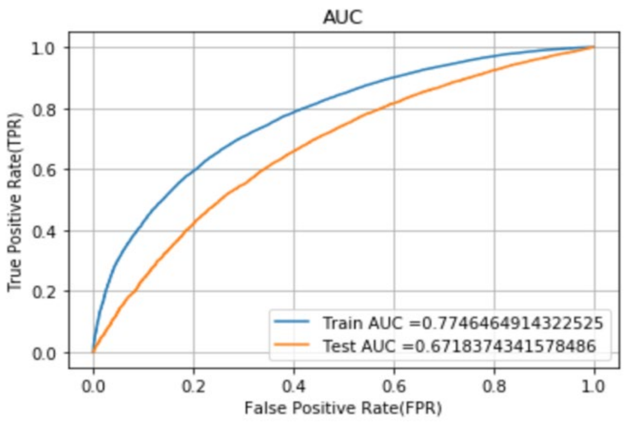
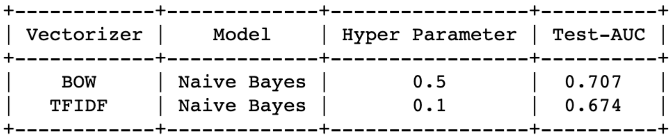


Figure 6. AUC of TF-IDF Naïve Bayes



```
+------------+-------------+-----------------+----------+
| Vectorizer |    Model    | Hyper Parameter | Test-AUC |
+------------+-------------+-----------------+----------+
|    BOW     | Naive Bayes |       0.5       |  0.707   |
|   TFIDF    | Naive Bayes |       0.1       |  0.674   |
+------------+-------------+-----------------+----------+
```

Figure 7. Comparison of Vectorizers

KNN, T-Singular Value Decomposition and random forest resulted in more than 60% accuracy. But this do not work so well with a huge number of data points and higher number of features. Distance based algorithms do not work well because the vectors will be grouped very close to each other that these models find it hard to separate different classes. Logistic regression, (see Figures 8, 9, and 10), Naïve Bayes (see Figures 5, 6, and 7) and neural networks (see Fig. 11) have shown promising results in improving the precision score above 70%.

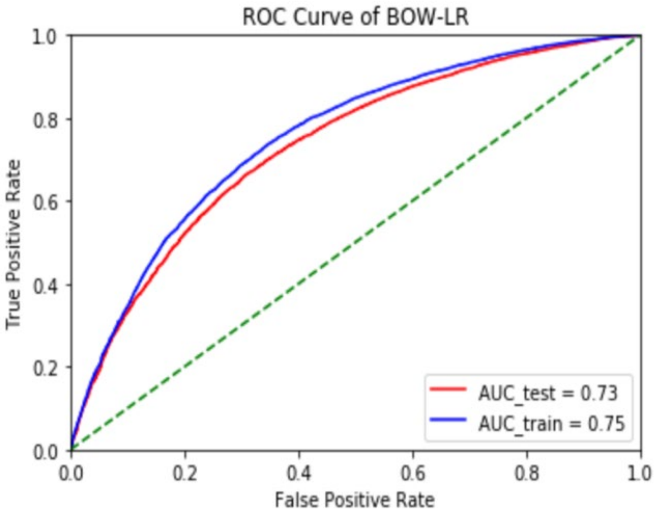


Figure 8. AUC of Bag of Word- Logistic Regression

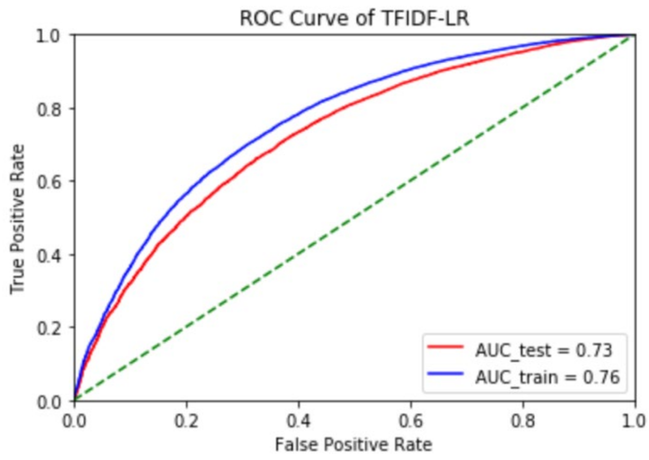
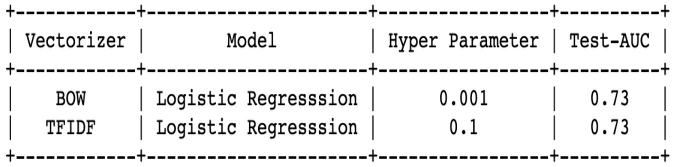


Figure 9. AUC of TFIDF- Logistic Regression



```
+------------+----------------------+-----------------+----------+
| Vectorizer |        Model         | Hyper Parameter | Test-AUC |
+------------+----------------------+-----------------+----------+
|    BOW     | Logistic Regresssion |      0.001      |   0.73   |
|   TFIDF    | Logistic Regresssion |       0.1       |   0.73   |
+------------+----------------------+-----------------+----------+
```

Figure 10. Comparison of Vectorizers

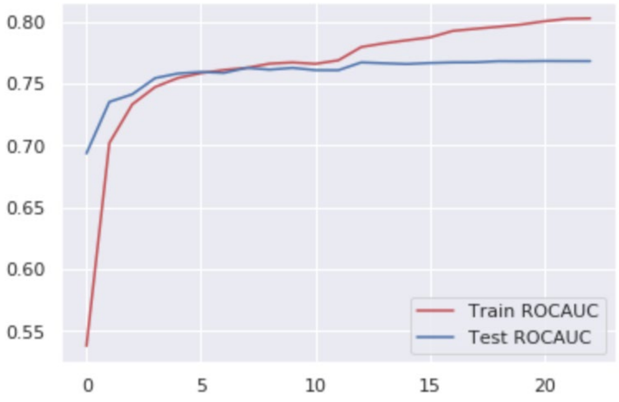
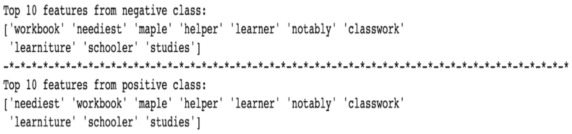


Figure 11. Neural networks

Sentiment Polarity is a sentiment value for a given English sentence; the score with 1 is very positive while 0 is very negative. We used this technique also to see if the model trains well. The top 10 feature from the negative can positive classes respectively are shown in Fig. 11.



```
Top 10 features from negative class:
['workbook' 'neediest' 'maple' 'helper' 'learner' 'notably' 'classwork'
 'learniture' 'schooler' 'studies']
-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
Top 10 features from positive class:
['neediest' 'workbook' 'maple' 'helper' 'learner' 'notably' 'classwork'
 'learniture' 'schooler' 'studies']
```

Figure 12. Top features:

V. Conclusion and Future Work

In this paper, we apply machine learning algorithms, and we analyze each feature of the Donorschoose dataset. Donorschoose is a readily available dataset which contains data about US school projects proposed by teachers. The proposed system will address the problem of retaining the existing donors by applying the machine learning models to analyze the behavior of the donors and filter the most promising projects from the dataset. The primary objective to apply machine learning is to extract 10 most important factors for a project to be picked by a donor. This will give a better idea to schoolteachers so that they can formulate their projects in more winnable ways. These features can also be crucial to the donors, to keep them donating to the schoolteachers.

From our analysis we noted that donors accept project with high level of project description because the essays seem to significantly impact the AUC value. Also features like project-resource-summary and datetime shown positive effect which states that projects with budget friendly and good explanation of the how the utilities are to be used are important, which is expected. Naïve bayes algorithm have extracted important words like workbook, helper, classwork as most words with high probability values in both negative and positive classes. It is also concluded that distance based algorithms suffer from curse of dimensionality when dealing with textual data and training the neural networks seems to be more promising.

References

[1] T. Althoff and J. Leskovec. Donor Retention in Online Crowdfunding Communities : A Case Study of DonorsChoose.org. Proc, ACM Press, 34–44, 2015.

[2] A. K. Agrawal, C. Catalini, and A. Goldfarb. The geography of crowdfunding. Technical report, National Bureau of Economic Research, 2011.

[3] J. Arguello, B. S. Butler, E. Joyce, R. Kraut, K. S. Ling, C. Rosé, and X. Wang. Talk to me: foundations for successful individual-group interactions in online communities. In SIGCHI, 2006

[4] M. Bichler and C. Kiss. A Comparison of Logistic Regression, k-Nearest Neighbor, and Decision Tree Induction for Campaign Management. AMCIS Proceedings. 230, 2014.

[5] C.R. Boyd, M.A. Tolson, and W.S. Copes. Evaluating trauma care: The TRISS method. Trauma Score and the Injury Severity Score. The Journal of Trauma. 27 (4): 370–378, 1987.

[6] D. Coomans and D.L. Massart. Alternative k-nearest neighbour rules in supervised pattern recognition : Part 1. k-Nearest neighbour classification by using alternative voting rules. Analytica Chimica Acta. 136, 15–27, 1982.

[7] P. Jaskowiak and R. Campello. Comparing Correlation Coefficients as Dissimilarity Measures for Cancer Classification in Gene Expression Data. Brazilian Symposium on Bioinformatics, 1–8, 2011.

[8] J. Pennington, R. Socher, and C.D. Manning. GloVe: Global vectors for word representation. EMNLP, 2014.

[9] J. Rennie, L. Shih, J. Teevan, and D. Karger. Tackling the poor assumptions of Naive Bayes classifiers. ICML. 2003

[10] F. Tom. An Introduction to ROC Analysis. Pattern Recognition Letters. 27 (8): 861–874, 2006.

[11] Donorschoose Dataset, available online at https://www.kaggle.com/donorschoose/io. Last accessed December 14, 2020.

[12] Donorschoose Dataset. Available online at https://www.donorschoose.org/data. Last accessed December 13, 2020.