Belen Michel Torino
DonorsChoose.org Project

Education Project Funds

Machine Learning classification for under-funded projects

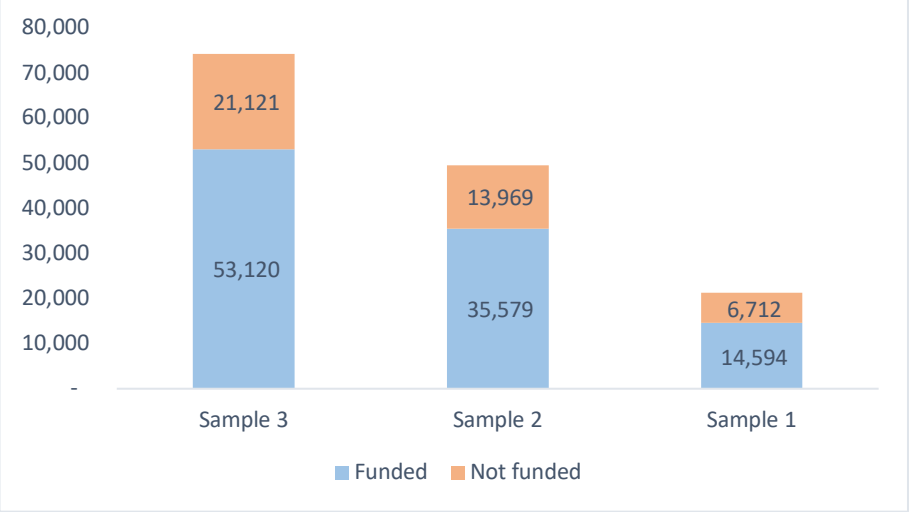
Summary Report

The present document will present the results of a study that seeks to predict if a project on the platform of DonorsChoose.org will not get fully funded within 60 days of posting.

DonorsChoose.org is an online charity that makes it easy to help students in need through school donations. At any time, thousands of teachers in K-12 schools propose projects requesting materials to enhance the education of their students. When a project reaches its funding goal, they ship the materials to the school.

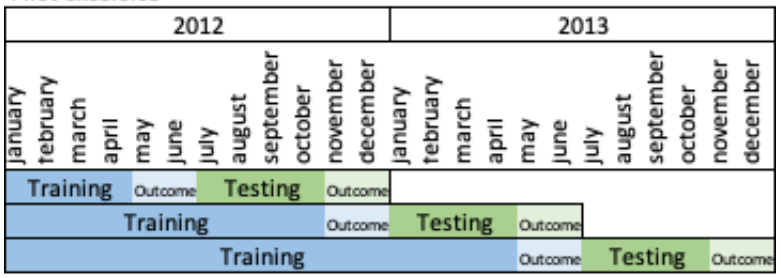
The present analysis will use different classification models to identify the projects that are not very exciting to the business and thus, prevents them from reaching full funding during the first 60 days. This prediction will be done at the time of posting so our analysis only use data available at that time.

The data that we used was in a file that has one row for each project posted. We identified our target projects as those that have more than 60 days between the date the project was posted ("date_posted") and the date the project was fully funded ("datefullyfunded"). As we can observe in the following graph we wanted to predict that orange portion of the projects that represent a approximately a 30% of all the projects. With this information we know that the baseline accuracy of our prediction model should be around 70%.

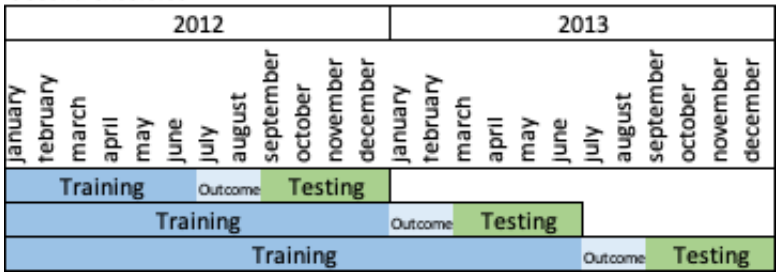


The data spans Jan 1, 2012 to Dec 31, 2013, thus we built 3 samples of data where our training sets start on the 1/1/12 and end 60 days before the testing set starts, allowing for the projects to be funded within those days. Moreover, we produced these samples twice, one with test sets extending over 6 month and another with testing sets extending 4 month -to account for cases where we would not know the funding's that occur after Dec 31, 2013. The following charts represent the different training and testing sets used:

First excersice



Second excersice



This study used the following Machin Learning models to predict the unfunded projects: a) Logistic Regression, b) Decision Tree, c) K-Nearest-Neighbors, d) Support Vector Machine, e) Random Forest Regressor, f) Ada Boost, g) Bagging.

Moreover, these models where tested with the following metrics: a) AUC_ROC, b) accuracy, c) precision, d) recall; at the following thresholds 1%, 2%, 5%, 10%, 20%, 30%, 50%.

The idea behind presenting different thresholds is that, if DonorsChoose.org wants to help some projects to be better developed and communicated, in order to achieve their objective fund faster then, they will be able to determine how many projects they are able to assist - represented by the percentage chosen as threshold- and thus, how many of the weaker projects they can expect to reach at that level. For example, if they are able to help 5% of the project they should know that with most of the models, 30% of the projects that they will be helping are going to be the weaker ones.

In order to be able to take the most relevant decision for their job, our analysis presents a table with the results where the organization can find the model that best suited their needs, in order to target the projects that have a higher risk of not being founded within 60 days.

Finally, if have to give a short recommendation without knowing the capacity and the exact objectives of DonorsChoose.org I would suggest them using:

A) The Logistic Regression model with the following parameters: penalty: l1 and C: 10.000; or
B) The Random Forest Regressor with the following parameters: n_estimators: 1, max_depth: 1, max_features: sqrt, min_samples_split: 2, n_jobs: -1.

With these models the AUC ROC is 0.646209 and 0.604108 respectively.