Short Report - Machine Learning analysis for DonorsChoose.org

Author: Luis Eduardo San Martin

Objective

The objective of the present report is to give recommendations to DonorsChoose to identify the upper 5% of school-related posted projects at the highest risk of not being fully funded after 60 days, so they can execute further actions with these projects. It also presents a brief summary of the analysis performed.

Analysis performed

The analysis consisted on the generation of several Machine Learning (ML) models to predict whether a school-related project will not get full funding after 60 days of having being posted. Its only data source was data from DonorsChoose.org, it consisted on data from 124,976 projects posted from January 1st 2012 until December 31st 2013. Prior to the ML analysis, the dataset was explored to find inconsistencies, duplicated values and missing data points. Then, a data preparation exercise was conducted, which consisted of discretizing continuous variables, generating dummies for categorical features, and imputing missing values.

Importantly, the preparation part also divided the whole dataset in three different sets using three time thresholds. For each threshold, the goal was to build a model with projects posted before the threshold and analyze its results on projects after the threshold. The first threshold was June 30th 2012, the second was December 31st 2012 and the third was June 30th 2013. The objective of this was to check if a good prediction model was consistent across different time spans, and that any good performance was not due to randomness.

After this, the ML analysis was performed. It used as predictors variables mostly related to the projects themselves (e.g: the subjects they are related to, the amount of funding they needed) and to the schools where they were to be conducted (e.g: their locations, the type of school). The analysis used different prediction methods to build a total number of 108 different prediction models, which were applied to each one of the three different temporal holdout datasets. Thus, a total of 324 different results were generated.

Results

The three results (one for each different temporal dataset) of each model were compared with the results of every other model using different performance metrics for prediction models, but with a focus on which models were better for predicting the projects at the highest 5% risk of not getting funding within 60 days of being posted. From now on, this report will call use the term “precision at the highest 5 percent” to refer to this metric.

In general, the results show that the best performing models across all the metrics evaluated are the ones that use Support Vector Machines (SVM) and Logistic Regression (LR) classifiers. It is

also important to notice that compared to other ML methods analyzed, the SVM and LR were computationally cheap.

Taking into account the three temporal holdouts used, the models tend to predict better the second and third temporal holdouts (the most recent ones, with thresholds on December 31 2012 and June 30 2013 respectively) than the first one. For example the best-performing model for the first holdout has a precision at the highest 5 percent of 45.4%, while the best-performing models in the second and third holdout have 58.7% and 52.1% respectively.

Analyzing each model and its results on precision at the highest 5 percent, the comparison does not show a single best-performing model for the three temporal holdouts. Nevertheless, a number of models consistently show up among the best ones in all our temporal holdouts. This report especially highlights the models with the following characteristics[1], in order to provide a final list with the best-performing ones.

- Models that use SVM or LR classifiers.
- Models that are amongst the best three for the second or third holdout, taking into account the precision at the highest 5 percent.
- Models whose precision at the highest 5 percent for the first holdout is at least 42.5%.
- Models whose precision at the highest 5 percent for the second holdout is at least 57%.
- Models whose precision at the highest 5 percent for the third holdout is at least 51%.

And those models are the following:

- LR with a penalty method of L1 and a C value of 1[2].
- LR with a penalty method of L1 and a C value of 10.
- LR with a penalty method of L2 and a C value of 1.
- LR with a penalty method of L2 and a C value of 0.01.
- SVM with a penalty method of L2, a primal optimization method and a C value of 0.001.
- SVM with a penalty method of L2, a primal optimization method and a C value of 0.01.

Recommendation

We recommend using any of the six highlighted models for the prediction objectives identified at the beginning of this report. If the audience of this report wanted a one-model-only recommendation, we recommend the LR model with a penalty method of L1 and a C value of 1 for being the best-performing model for precision at the highest 5 percent in the most recent temporal holdout (the third one).

[1] This characteristics do not constitute whatsoever a benchmark on what defines a good model on any ML analysis. They were selected given the particular characteristics of this analysis and its results: (i) the performance metrics of the models vary significantly across temporal holdouts, (ii) SVM and LR models tend to perform better than any of the other types analyzed, and (iii) the audience for this report is especially interested in identifying the projects at the highest 5% risk of not getting funding.

[2] The rest of the model parameters in all of these cases are the ones defined by default using the sklearn package from Python3.