

Predicting academic-challenge success

Dante López, Luis Villaseñor, Manuel Montes-y-Gómez,
Eduardo Morales and Hugo Jair Escalante

Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)
Sta. María Tonantzintla, CP 72840, Puebla, México.
{danterss, villasen, emorales, hugojair}@inaoep.mx

Abstract. Academic competitions and challenges comprise an effective mechanism for rapidly advancing the state of the art in diverse research fields and for solving practical problems arising in industry. In fact, academic competitions are increasingly becoming an essential component of academic events, like conferences. With the proliferation of challenges, it is becoming more and more relevant to distinguish potentially successful challenges before they are launched. This in order to better allocate resources, time slots, sponsorship and even to have a better estimate of expected participation. This paper presents a first study in this direction: we collected a data set from Kaggle and aim to predict challenge success by using information that is available before a competition starts. We characterize competition proposals by textual information and meta-features derived from information provided by organizers, and use these features to predict challenge success (estimated by the number of participants and submissions). We show that both, text and meta-features convey predictive information that can be used to estimate the success of an academic challenge.

Keywords: Prediction of challenge success; Academic competitions; Challenge proposals

1 Introduction

Academic challenges dealing with data analysis represent an increasingly popular means for solving applied problems and advancing the state of the art in a number of fields. Challenge organizers set a relevant problem, and provide data, an evaluation protocol, rules and incentives. Then, the crowd aims to solve the problem by making use of every skill and resource, provided they do not harm challenge rules. Challenge campaigns and evaluation forums have a long history within academia (e.g., DARPA programs¹, TREC[1] and ChaLearn challenges [2]) and industry (e.g., the Netflix prize [3]) and together have pushed the boundaries of science and succeeding in practical solutions relevant to industry/society problems.

With the availability of massive amounts of heterogeneous data being generated constantly, data analysis tasks have diversified and increased their reach in terms of societal and economical impact. This has caused a proliferation of challenges being organized and running at the same time. Since challenges are often collocated with academic events (e.g., conferences, workshops, etc.), proposing an academic challenge is getting more and more competitive: many challenge proposals and limited resources (e.g., time slots, sponsorship and even participants). Because of this, automatic ways of determining the success of challenge proposals would be extremely helpful for conference chairs, sponsors and challenge organizers themselves.

¹ <https://www.darpa.mil/our-research>

This paper explores the feasibility of predicting challenge success from information available before a challenge starts (e.g, from a challenge proposal). We build a novel data set with challenge proposals taken from Kaggle² (the most popular platform for challenge organization) with the aim of exploring the feasibility of the task. We define a set of meta-features that combined with textual information are used as predictive features for estimating the success of a challenge (measured as the number of participants and submissions). We analyze the performance of regressors and classifiers and show that the features convey useful information that allows us to predict the variables of interest. In addition, we analyze the most discriminative features, which can give insights for organizers. To the best of our knowledge this is the first effort in such direction.

The remainder of this paper is organized as follows. Section 2 describes the collected data set. Section 3 formulates the problem and introduces the meta-features and textual information considered for our study, also we describe the adopted supervised learning framework. Section 4 reports experimental results. Finally, Section 5 outlines conclusions and future work directions.

2 The Kaggle data set

For building the associated data set we relied on the most popular platform for challenge hosting: Kaggle. It was founded in 2010 and it was recently acquired by Google. The focus of Kaggle is on predictive analytics challenges. We postpone for future work experiments and evaluation on other popular platforms (e.g., CodaLab³).

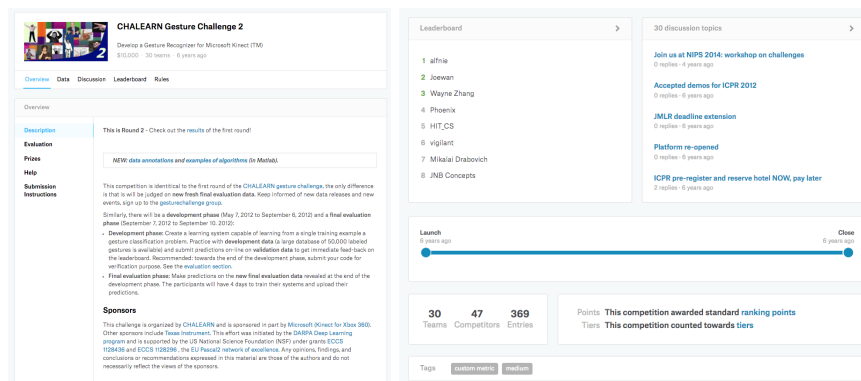


Fig. 1. Snapshot of a typical Kaggle challenge.

We considered all of the finished competitions up to April first 2018. Completed competitions were required to build the corpus because we needed to have a way to determine the extend of success of each competition. In total 270 of Kaggle competitions were considered for the study. Figure 1 shows a snapshot of a typical Kaggle challenge website. We aimed to collect for each competition most of the information available, with emphasis on potentially discriminative information. Textual information was gathered form the overview description, evaluation, and data sections. Additional

² <http://kaggle.com/>

³ <http://codalab.org/>

information comprising duration, prizes, number of registered teams, and tags was also compiled. Finally, we also stored information on the scores from the top ranked participants. The next section explains the set of meta-features extracted from the corpus to be used by the predictive models and the criteria we used to determine whether a challenge was successful or not.

3 Predicting challenge success

We focus in the problem of predicting a variable related with the success of challenges starting from information derived available before the challenge starts. Success of challenges can be measured in a number of ways: e.g., by determining how close the winning solution was from the optimal performance, by determining whether the winning solution was acquired by the organizer (e.g., a company), by the number of participants or teams involved, by the average performance of participants etc. Among these options, we opted to associate the success of a challenge with the number of registered teams and the number of submissions made by the top ranked participants. Our choices are based on the intuition that:

1. *A challenge is successful if it is attractive to people (measured by the number of participants);*
2. *The number of submissions by top-ranked teams reflects the engagement of participants in the competition (measured by number of submissions among the top ranked participants).*

Therefore, we aim to estimate both of these variables that are directly related to challenge success. Please note that we do not opt for predicting the "successful vs. non-successful" variable⁴ because it may be arguable how many participants/submissions make a competition successful. Instead we aim to estimate the number of participants and submissions, which can be a more useful variable (i.e., in addition to measure the success of a challenge, these predictions can be helpful to allocate resources for competitions).

3.1 Features

The features that were used to build predictors were extracted from information that is available before a competition starts. This is motivated by the fact that challenge organizers often prepare challenge proposals that are analyzed, judged and selected by conference chairs and sponsors (see e.g., [4]). We think our model could be used by people that decide on the collocation/sponsorship of challenges as an additional support tool. Hence, we used information that is commonly available in a challenge proposal. The set of derived features is presented in Table 1.

We distinguish two types of features/information: textual features, extracted from the text in the challenge site (see Figure 1) and features derived from the challenge website (we called these meta-features to make clear the distinction from textual features). The 14 considered meta-features capture information that is, intuitively, engaging and could be related with challenge success. For instance, the amount of offered prize, the inclusion of images in the description, the number of allowed submissions, models, etc. On the other hand, textual information was considered because we wanted to study whether what organizers say may be correlated with the success of the challenge. Text was represented using a bag of words with TFIDF weighting scheme. Stop words and punctuation marks were removed, all text was converted to lower case before indexing. We performed experiments using the whole corpus vocabulary and reducing the number of dimensions by using mutual information [5] (below, the best results are described). As described below, we performed extensive experiments merging feature representations and normalizing the outputs.

⁴ Although we report some results by discretizing the output variables, see Section 4.

Table 1. Features considered in this study.

Name	Type	Description
Meta-features		
Category	Type of competition: research, recruitment, novices, experts	Categorical
Tags	Associated keywords	Categorical
Type	Type of submission: code, prediction, recreation	Categorical
Prize	Amount of offered prize	Numerical
Currency	Currency of offered prize: USD, Euro, Other	Categorical
Images	Binary variable: is an image included in the description?	Binary
Duration	Challenge duration	Numerical
Data type	Text, image, numerical, other	Categorical
Winners	Number of offered prizes	Numerical
Score	Score obtained by the baseline of the competition	Numerical
Metric	Evaluation measure: RMSE, Accuracy, Log loss, etc.	Categorical
Members	Maximum number of participants per team allowed	Numerical
Submissions	Maximum number of submissions per day	Numerical
Solutions	Maximum number of final models per team	Numerical
Textual-features		
Description-text	All text used in the description of the competition	Bag of words
Evaluation-text	All text used in the description of the evaluation section	Bag of words
Data-text	All text used in the description of the Data	Bag of words

3.2 Predictive modeling

As previously mentioned, we approached the problem of predicting the number of teams registered to a competition and the average number of submission made by the top ranked participants (top-10 were used). Accordingly, popular regression methods were considered to approach the problem. In preliminary experiments we discarded several models because they were showing very low performance (e.g., linear regression, Gaussian processes and support vector regression). At the end we used random forest regression as the predictive model. Implementations in the popular Sci-kit learn library were used for all of our experiments [6].

In addition to addressing the regression problem, we formulated the problem as one of classification. Our aim was to study the feasibility of the task when trying to estimate categories instead of raw values. Artificial classes were defined by dividing the numerical outputs into three categories for each predictive variable. The generation of intervals was done manually and aiming to keep balanced classes. The induced artificial classes are shown in Table 2.

Please note that the generation of these artificial classes was simply done to study whether classifiers perform better than regressors. Our goal was not to determine thresholds on the variables to distinguish successful challenges. For approaching the classification problem we considered the following classifiers: support vector machines, naïve Bayes, and j48 decision tree.

Table 2. Artificial classes per variable. The number of samples in each class is shown between parentheses.

Variable	Class 1	Class 2	Class 3
Number of participants	≤ 200 (97)	$> 200, \leq 700$ (88)	> 700 (85)
Number of submissions	≤ 30 (114)	$> 30, \leq 70$ (83)	> 70 (73)

4 Experiments and results

This section reports experimental results on the estimation of challenge success from competition proposals. As previously mentioned the generated corpus comprises 270 samples, for experimentation we adopted a cross validation strategy: we report average 5-fold cross validation performance (the same folds were used for every experiment reported herein). Regression performance was evaluated with the root-mean-squared-error (RMSE, the lower the better) and the determination coefficient (R^2 the higher the better). Classification performance was evaluated with the average of f_1 measure across classes.

4.1 Challenge success prediction as regression

This section reports experimental results when predicting the number of participants and submissions directly. The problem is approached as one of regression. We tried two configurations, in the first one we predict raw values as extracted from the data set. In a second experiment we normalize the outputs to the range $[0, 1]$ and aim to predict this normalized outputs. We evaluated normalized outputs because the ranges of values in the raw values were rather large, which caused the regressors to be unstable for some samples (see below). Also, intuitively having a predictor that provides estimates as a bounded real number would be more helpful if one want to assess the successfulness of challenges in relative terms (the perfect challenge would get 1 and the worst 0). Table 3 shows the performance obtained when using meta-features, textual information and a combination of both types of features (early fusion scheme), and Figure 2 shows plots of ground truth values vs. predictions for the different configurations.

Table 3. Challenge success prediction performance.

Configuration	Raw		Normalized	
	RMSE	R^2	RMSE	R^2
Participants				
Meta-features	1005.41	0.11	0.10	0.12
Text-features	939.36	0.27	0.09	0.29
Meta-Text-features	923.48	0.29	0.09	0.32
Submissions				
Meta-features	51.81	0.21	0.21	-0.27
Text-features	55.71	0.07	0.19	0.07
Meta-Text-features	51.94	0.19	0.19	0.05

From Table 3 it can be seen that error in raw estimations is somewhat large, mostly when estimating the number of submissions. This is decreased when estimating the normalized values. Large error values are partly due to the great variability in the range of values that such variable can take. This can be graphically seen in the left plot of Figure 2, where an outlier is affecting the performance of all methods. From the same plot, it can be seen that, in general, the predictions from the three regression methods are rather conservative (they underestimate the number of participants). Interestingly, when predicting the number of participants, textual information reported better performance than meta-features, and the combination obtained the best overall performance. In fact, the improvement offered by textual features is considerable: about 9% in absolute RMSE and much more in terms of R^2 score. This is a very interesting result, as it suggests that the words that are used by challenge organizers to describe their competition indeed affect the number of potential participants.

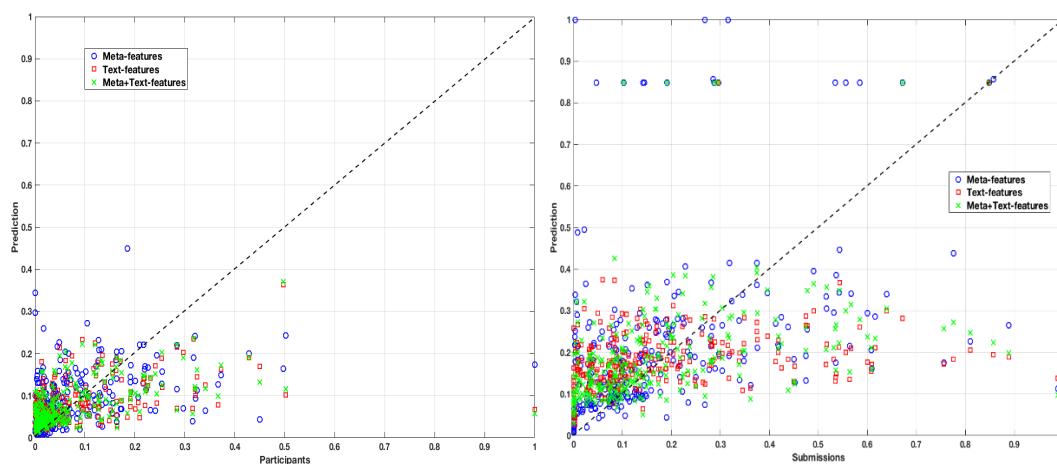


Fig. 2. Ground truth vs. predicted values. Left: predicting the number of participants. Right: predicting the number of submissions.

Regarding the prediction of the number of submissions it can be seen from Table 3 that the best performance is obtained when using meta-features only. This time adding text does not help to improve the performance. Graphically, from the right plot of Figure 2, the predictions using meta-features are slightly closer to the ideal line (although optimistic/large values are predicted for some challenges). The variants including text are more spread and this causes a larger error.

4.2 Challenge success prediction as classification

As previously mentioned, we also performed experiments by generating an artificial classification problem by thresholding raw values on the number of participants and submissions. Our hypothesis was that by discretizing the outputs would lead to satisfactory performance when predicting challenge success. Experimental results are shown in Table 4.

From this table it can be seen that performance is low in general, in fact, several configurations perform lower than a trivial baseline like predicting the majority class (35% and 42% for participants

Table 4. Challenge success prediction performance as classification.

Configuration	SVM	N. Bayes	j48
Participants			
Meta-features	0.33	0.27	0.55
Text-features	0.17	0.46	0.49
Meta-Text-features	0.43	0.22	0.52
Submissions			
Meta-features	0.25	0.26	0.49
Text-features	0.20	0.40	0.40
Meta-Text-features	0.51	0.26	0.5

and submissions, respectively). Among the considered classifiers, the best performance was obtained by the decision tree. In terms of features, both performed similarly, not clearly showing a tendency. Results from this experiment seem to indicate that the way we discretized the outputs may not be the best, in addition, since the range of values of the outputs are large, approaching the problem as one of classification may not be a good choice. Yet, being optimistic we think the results are promising, further research is needed in order to obtain satisfactory results.

Figure 3 shows the first nodes of the decision tree when using meta features for predicting participants (top) and submissions (bottom). It is somewhat expected that the root node is the *prize* feature. Interestingly, for discriminating class 2/3 samples (high number of participants), it is important if the competition category is either text or images; whereas a small number of submissions is mostly associated with class 1 samples. Regarding the number of submissions, the performance of the baseline causes a low number of submissions (class 1), whereas, a small number of daily submissions is associated with challenges having a large number of submissions.

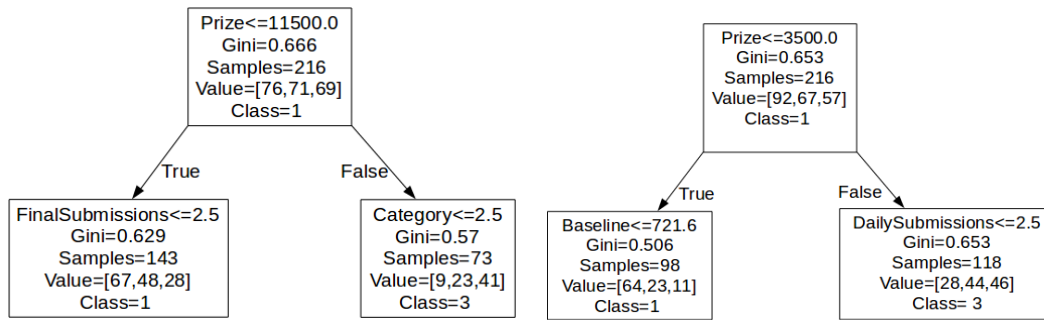


Fig. 3. Snapshot of the root of decision trees when using meta-features for predicting participants (left) and submissions (right).

Regarding textual features the most discriminative words for high participation were: *faq*, *mode*, *prepared*, *mitigate*, *properly*, *organization*, *effects*, while for low participation the following words were more important: *depending*, *enforce*, *shall*, *not*, *breaks*. We leave the reader to draw their own conclusions regarding the importance of this features. Finally, when mixed features were considered,

words dominated the earlier levels of the tree for both variables participants and submissions. Interestingly, new words emerged that were associated to class 3 like: Alex, Deep, Deeper, etc.

5 Conclusions

Academic competitions are becoming increasingly popular among scientist and practitioners of data science. With this overwhelming number of challenges being organized it is becoming more and more relevant to anticipate the success of challenges, so that organizers can allocate resources, sponsorship, and number of potential participants attending academic venues. This paper comprises a first step in this direction: we introduced a novel data set comprising both textual and meta features, and deployed and analyzed predictive models for estimating challenge success (via number of participants and submissions). The following conclusions are derived from this work:

- Experimental results reveal that the considered features convey useful information for predicting challenge success in terms of the two considered variables.
- Textual information is more more helpful for estimating the number of participants than meta-features, whereas the latter were more helpful for predicting the number of submissions.
- Approaching the problem as one of classification shows the problem is quite complicated and that the performed discretization may not be the best.
- The analysis of features revealed that the *prize* meta feature was the most discriminative, and among text, words with negative connotation (e.g., *not*, *enforce*, *breaks*, *depending*, *etc.*) were negatively correlated with challenge success prediction.

Future work includes extending the data set by including information from other competition platforms, and performing a complete linguistic analysis of textual features.

References

1. E. M. Voorhees and D. K. Harman. *TREC: Experiment and evaluation in information retrieval*. MIT Press, 2005.
2. Sergio Escalera, Xavier Baró, Hugo Jair Escalante, and Isabelle Guyon. Chalearn looking at people: A review of events and resources. In *Int. Joint Conference on Neural Networks*, pages 1594–1601, 2017.
3. Robert M. Bell and Yehuda Koren. Lessons from the netflix prize challenge. *SIGKDD Explor. Newsl.*, 9(2):75–79, December 2007.
4. ChaLearn. Chalearn procedures for organizing events and awarding prizes and travel grants. <http://chalearn.org>, 2013.
5. Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *ICML*, pages 412–420, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.
6. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *JMLR*, 12:2825–2830, 2011.