Name: Yoon Kim, Peng Liu
Location: New York, NY, USA
Competition: KDD Cup 2014

1. Summary

We worked independently and ended up just averaging our models in the end. Our features ended up being quite similar, though. In general it was important to have out-of-time validation sets. One of us used an external data source from ESLI, which gave very slight improvements in performance. Instructions to download the dataset is in section 6.

2. Features Selection / Extraction

Here were some potentially interesting features that we utilized for each model.

Model 1 (ensemble of 4 GBM models plus 1 ExtraTrees model)

- "History" variables (e.g. How many is_exciting projects did this school have in the past? How many great_chat projects did this teacher have in the past? How many unique donors in this zip code? etc.)
- We found that it was better to use (credibility-adjusted) historical rates for some categorical variables (e.g. historical is_exciting rate for teacher, pulled towards population mean, as opposed to actual counts).
- Using history features for other responses (great_chat, fully_funded, one_non_teacher_referred_donor_giving_100_plus, etc.) also gave us some decent gains.
- To account for time trend, we had features called avg_X_prediction where X was a sliding window of biweekly/monthly/bimonthly predictions from an initial model.
- For text, we created features by using logistic regression with tf-idf on title/essay using a leave-10%-out scheme to make sure that the response variables were not used twice. Predictions from this model were used as features in a GBM.
- Historical variables were calculated April 2010 onwards. The actual models were built on 2011+ data.

Model 2 (ensemble of GBM, ExtraTrees, Random Forests, and Elastic Net)

- Instead of using history variables, we used leave-one-out credibility adjusted (with jitter) rates.
- We also got very little performance boost from using publicly available ELSI data (http://nces.ed.gov/ccd/elsi/tableGenerator.aspx)

3. Modeling Techniques and Training

Hyperparameters were tuned via some light grid-search on a validation set of 2013-09-01 to 2013-12-31 for model 1, and 2013-06-01 to 2013-12-31 for model 2.
For GBM we tuned for (i) number of trees (ii) interaction depth (iii) minimum observation in node. Bagging fraction was set to 0.5 for all models. Shrinkage was set to 0.1 for GBMs in model 1 and 0.01 for GBM in model 2.

For Extra Trees/Random Forests we tuned for number of trees and number of variables considered for split at each node. For glmnet we tuned for alpha parameter (we set lambda = 0—i.e. we used ridge regression).

Both model 1 and model 2 are ensembles of other models. Ensembling weights were based on trial-and-error. Invariably the GBM models were weighted heavily.

Final model was simply an average of model1 and model2 (with discounting applied after averaging). We used a simple linear decay from 1.0 to 0.5 from 2014-01-01 to 2014-05-45.

4. Code Description

Code is available at:
https://github.com/yoonkim/kdd_2014

5. Dependencies

The following software/packages were used for the competition

Python – nltk, sklearn, numpy, pandas
R – gbm, extraTrees, sqldf, plyr, dplyr, randomForest, glmnet

6. How To Generate the Solution (aka README file)

Let -path be the folder with all the data.

To run:
1. Run "python kdd_2014_data_model1.py -path" in command prompt
2. Run kdd_2014_model1.R (change "folder <- -path" to appropriate path)
3. download ESLI data from http://nces.ed.gov/ccd/elsi/tableGenerator.aspx
(Public School, Years 2011-2012, columns school id and school type)
4. Run kdd_2014_model2.R (change "folder <- -path" to appropriate path) until line 126.
5. Run "python kdd_2014_data_model2.py –path" in command prompt
6. Final prediction is
(0.5*model1+0.5*model2)*discount

7. Additional Comments and Observations

- Our final submissions included one with linear discounting from 1.0 to 0.5, and one without any discount. The submission with no discount would have obtained 5th place on the private LB.

- Given private LB's sensitivity to discounting, and given public LB's (relative) lack of sensitivity to discounting (e.g. 1.0 to 0.5 linear decay gave ~0.003 improvements on the public LB), we were simply lucky.

- In order to emulate the LB we briefly experimented with weighting schemes (e.g. weighting is_exciting projects that were funded in 2 weeks more than those that were funded in 3 months) and censoring (e.g. only counting is_exciting projects if funded in one month). But these didn't affect things much (on the public LB).

8. Simple Features and Methods

A single GBM with history variables combined with discounting would have comfortably produced a model in the top 10.

9. Figures

NA

10. References

NA