ANLY 533 - Data Mining Module Assignment - Machine Learning for Classification

Kaggle Porto Seguro Insurance Challenge - Predicting Who Will File a Claim

Due Date: November 22, 2017 @ 11:55 PM

Assignment Description

You will be "competing" to build a model that can be used to predict whether or not a driver insured with Porto Seguro will file an insurance claim. Basic instructions for the Kaggle challenge, including training and test data sets, can be found here: https://www.kaggle.com/c/porto-seguro-safe-driver-prediction. After downloading and unzipping the training.csv and test.csv files, you will use the training set to build and validate a model that can predict whether or not a driver will file a claim.

The training and test files contain the same predictors. The only difference is that the test set does not contain target labels. The Kaggle competition provides the following information about the anonymized features:

- In the train and test data, features that belong to similar groupings are tagged as such in the feature names (e.g., ind, reg, car, calc).
- In addition, feature names include the postfix bin to indicate binary features and cat to indicate categorical features.
- Features without these designations are either continuous or ordinal. Values of -1 indicate that the feature was missing from the observation.
- The target column signifies whether or not a claim was filed for that policy holder.

Your task is to use the training set in order to predict results for the holdout test set. After selecting your final model, you will generate predictions for the test set and save the predictions to a file that is formatted exactly like the sample_submission.csv file contained in the competition zip download. After saving the CSV containing your predictions, you will upload your predictions to Kaggle and submit them to the competition.

Deliverables

This assignment allows you to demonstrate a variety of relevant modeling-related tasks, including data prep, feature selection, model selection, and tuning. The modeling approaches you use for this project are totally up to you, but for the final product, you need to provide:

You are required to submit 4 items:

1. A set of binary scores (0 or 1) that are predictions for the entire test set, as a CSV with one column for the anonymized driver ID (id) and a second column containing your predicted value for each driver (0 or 1).
2. The code you used to produce the predicted values with documentation (in-line comments and/or an R Markdown file or Word Document) outlining the thinking behind the choices you made and steps taken in order to produce the final scores and the diagnostic plots you made to evaluate your final model(s)
3. A brief (1-2 page) write up describing

- The business problem (Why would we want to predict who will file a claim?)
- The data (What data is available and how can it be leveraged?)
- Data preparation (How did you clean the data to prep it for modeling? If you created new features, how did you decide what to do in that stage of the process?)
- Modeling (Variable selection, model evaluation/interpretation, cross valiation)

- Evaluation (Have you solved the business problem sufficiently? Why or why not?)
- Deployment (Does your final model make good predictions?)

4. Proof that you submitted your test set to the kaggle competition and that the file was accepted/scored (screenshot confirmation is fine).

Specifics

You need to try at least 3 different machine learning techniques for classification that we have discussed in class

You may also try other classification methods if you would like.

After training the various classifiers, you need to use evaluation methods from class to decide which one is best, and then continue tuning as appropriate to build the final model you decide to use to generate your predictions.

Grading/Evaluation of Your Model

Your performance will be evaluated less on the predictive accuracy of the final scores, and more on the process used to generate them. To this end, here are a few suggestions:

- You have a limited amount of time to complete the assignment, and training classifiers on complicated data sets can be time consuming, so plan accordingly. **You cannot wait until two days before this assignment is due to begin training models**
- An adequate set of finished predictions is highly preferred over predictions which look very promising but aren't completed by the deadline. The scores' accuracy is a part of the grading process, but it is far more important to see how you produce the scores.
- At a minimum, your prediction scores should be at least slightly better than a baseline model that predicts the majority class for every case.
- Documentation can be organized in the code or separately, but either way, I should be able to follow along with your reasoning throughout the model building and evaluation process.
- Don't get bogged down and spend tons of time on feature engineering or on training tons of different models (I understand that you are all part-time students and that you have other things to work on at this point in the semester), but you should demonstrate that you understand these processes.
- **You are all novice modelers...** A thoughtful write-up/documentation is much preferred over a very good model. I want to know that you are able to organize your workflow methodically through the modeling process to produce a well executed set of predictions. **I want to be able to clearly see how you make decisions through the process.** It is not very important that you achieve a high score on the Kaggle leaderboard.