Stat 222 Project 4: Kaggle Competition

1 Data Description

Kaggle is a website that hosts predictive modeling competitions. Some of you may already have participated in these competitions on your own or in other classes. For this project, your group needs to

- choose a competition
- familiarize yourself with the data and the problem context
- brainstorm & research methods to implement in solving the prediction problem
- implement your methods and make at least two submissions to Kaggle
- document your work in a technical report.

You do not need to choose an active competition, but you may find it more fun to do so. One important limitation is that *you should choose a competition that none of the group members has worked on previously.* It is ok if more than one group chooses the same competition.

You will need to sign up for an account with Kaggle if you have not already. When you submit your predictions, do so as a team. For more information on this, see https://www.kaggle.com/wiki/FormingATeam.

2 Project Proposal: Due Friday, March 20

Since this is the most open-ended project we've done so far, it's important that your plan has the appropriate scope. To help you achieve this, we'll be giving you some early feedback about your project. By 5pm Friday, March 20, write a 1-2 page proposal describing what you plan to do and email it to both the instructor and GSI. This does not need to be formal, but you should NOT cut and paste any wording from the project website. You do the preliminary work of reading the documentation, downloading the data, and doing some exploratory data analysis, so can address each of the following points in your own words.

1. Competition name and link
2. Data description - What are the variables? Describe any relevant characteristics (dimension, distribution, temporal indexing, etc.)
3. Prediction objective - What do you need to predict? How will predictions be evaluated? (Give a mathematical expression if possible.)
4. Potential methods - What methods will you try? Why are they appropriate?
5. Relevant benchmarks - What simple methods (e.g. random guessing, deterministic solutions, or off-the-shelf algorithms) can you compare your results to? Often you can find these in the leaderboard.
6. Helpful references - What books, articles, or other resources can you consult in learning about either the subject-matter of the competition or the methods you plan to use?
7. Potential pitfalls - What do you anticipate will be difficult about this prediction problem?

3 Project Updates: Due Wednesday, April 10

Each group should plan to give a 10 minute overview of your progress in class on April 1. You may use slides if you like, or just talk informally. I expect you will have preliminary plots and/or results to share. The goal here is for the class to be exposed to the variety of problems that are being worked on, as well as for you to solicit feedback and ideas from people outside your group.

4 Project Timeline

- Wed, March 18 - project intro; work day
- Fri, March 20 - proposal due via email
- week of March 23 - spring break
- Mon, March 30 - work day
- Wed, April 1 - project updates in class
- Mon, April 6 & Wed, April 8 - industry speakers in class
- Fri, April 10 - tech report due via email