# Stat 502 Spring 2018 Project Statement (V1.0)

## Generalities

Every Stat 502 student must complete (as an individual or fully participating member of a project group) a large predictive analytics/supervised learning project. The problems that may be used in this project are:

1. The Data Mining Cup (DMC) competition sponsored annually by Prudsys AG. Details can be found here: https://www.data-mining-cup.com/
2. The Kaggle TalkingData AdTracking Fraud Detection Challenge (a presently running open Kaggle competition). Details can be found here: https://www.kaggle.com/c/talkingdata-adtracking-fraud-detection
3. The Kaggle ("In Class" private) House Price competition (created by Vardeman and running until 11:59PM April 27). Details can be found here: https://www.kaggle.com/c/isu-stat-502-2018-a

These projects are listed above in order of decreasing size, likely complexity, and likely time investment required in order to be reasonably successful. The first is traditionally organized into a single ISU effort ultimately split into 2 submissions to Prudsys and will almost certainly involve team members outside of the present Stat 502 class. The number of 502 class members joining the DMC competition is not limited, but those choosing to join must be aware of the very high time and effort commitment that is required in this competition. The second is what looks like a pretty big (by Kaggle standards) problem, and may well be unwieldy-to-impossible to do without using either Kaggle's computing platform or a cluster. For Stat 502 purposes, groups of up to size 4 may be entered in this second competition. (Not all group members need to be 502 students, but for our purposes, the limit on total group size will be 4 unless otherwise specifically authorized by Vardeman.) The third is a comparatively small and very well-formulated prediction problem. Much (if not all) of it can be done on one's own reasonably capable computer. For Stat 502 purposes, individuals or pairs of individuals may be entered into this competition. (There is no real reason for someone outside Stat 502 to have competitive interest in this problem, so we'll limit team membership for 502 purposes to the class. If this restriction or the restriction to singles or pairs is for some good reason a serious issue for you, you may inquire with Vardeman about a small relaxation of these restrictions.)

While everyone in Stat 502 must turn in HW#3, additional HWs for 502 will be optional/"Extra Credit" for DMC and TalkingData participants, but  be required for those taking the House Price option (effectively becoming part of a hybrid "final project" consisting of both a competition problem and some HW not absolutely required of others).

Ultimately, there is more to be learned from a DMC or TalkingData effort than from the House Price competition and I don't want to dissuade those of you who can afford to work on one of those from doing so. But I also don't want this project to be an impossible burden, hence the option of the much smaller House Price Prediction problem.

## Sign-up

By the end of the day Friday March 23, each group must send Vardeman a single e-mail stating:

1. The option/problem chosen.
2. A list of all team members (including both those taking Stat 502 and any not taking Stat 502).
3. In cases where option #2 or #3 is chosen, the official Kaggle name of the team (so Vardeman can track progress on the leaderboards if he wishes to do so).

## Reporting Requirements

### "Interim" Reporting

There are two templates presently on Canvas for weekly reporting of all team activity and progress. Beginning with the week of March 26 (if not March 19) these are to be filled out for the week by class time on Fridays. Those choosing option #2 or #3 have two templates to fill out as the course goes on, as they will have scores from the Kaggle Public Leaderboard to help guide their efforts and one template is for reporting progress there.

All teams will need to add the week to the "Weekly Updates Template." For each week, if appropriate, begin with a sentence or two describing overall activity, decisions, etc. Then make a bullet for each team member and describe in a sentence or two the person's activity for the week (beginning and ending at class time on Friday).

Depending upon how things go, we may spend some class time on Fridays reviewing progress and discussing issues of potentially general interest to the class. I will try to see if I can figure out how to let you post the updates to the interim reporting documents directly to Canvas. (Right now, I don't know what's doable.)

### Final Reporting

Each team with on-campus members will make an oral report on its work during the regularly scheduled Final Exam period, Monday April 30. Wholly off-campus teams will need to put together a video or "voice-over-powerpoint" presentation that can be viewed in that same Final Exam period (and thus made available to all).

Each team will submit a single written final report on its work.

1. The Stat 502 part of the DMC group will need to prepare a full-blown technical description of the problem set by Prudsys, their approach and work through Friday April 27 (including technical details as they are needed), and a clear statement of what else is planned before the contest deadline. The maximum length of this paper should be 20 pages plus any appendices (1.5 line spacing with 11 point font in .pdf format). Some more guidance regarding what needs to be addressed will be forthcoming.
2. Teams choosing Kaggle competitions will need to prepare a less extensive (but clear and accurate) description of their work. The maximum length of this paper should be 10 pages plus any appendices (1.5 line spacing with 11 point font in .pdf format). Here too, some more guidance regarding what needs to be addressed will be forthcoming.

Each person in a group of two or more will send Vardeman an e-mail with the subject line:

Team XXXX Peer Evaluations by YYYY

and in that email list every Stat 502 team member and give each person (**including themselves**) "Effort" and "Value" ratings. (The intent is to divide up both the "amount" of work done by the team and the intellectual contributions of team members to its progress.) These ratings must add to 100 across team members (both for Effort and separately for Value). The format for the body of that email is (listing members alphabetically by last name):

| | Effort | Value |
|---|---|---|
| AAA | 25 | 30 |
| BBB | 20 | 50 |
| CCC | 30 | 10 |
| DDD | 25 | 10 |

If explanatory comments are desired, they may be provided after this peer scoring.

## Instructor Evaluation

Vardeman will be faced with the unenviable task of informally processing the information in the interim and final reports and try to arrive at some evaluation (grade) for Stat 502 project work.