Stat 27850/30850: Group project # 2 (Data option)

Data The data set for the first project is the Kepler Exoplanet Search data set, available from `https://www.kaggle.com/nasa/kepler-exoplanet-search-results` (see also `https://exoplanetarchive.ipac.caltech.edu/docs/API_kepcandidate_columns.html` for more details about the variables in the data set and other information). This data comes from the Kepler Space Observatory telescope. The telescope identifies "Kepler objects of interest" (KOIs), which are observations from the telescope that potentially correspond to an exoplanet. These potential exoplanets are then classified as "CONFIRMED" (i.e., a true positive) or "FALSE POSITIVE", or (if not yet classified) then it's listed as "CANDIDATE".

On Canvas, we provide a slightly cleaned data set `KOI.csv` containing 9564 KOIs. The variables are:

- `kepid`, the unique ID for this observation.
- `koi_disposition`, one of "CONFIRMED" or "FALSE POSITIVE" or "CANDIDATE".
- `ra` and `dec`, which indicate the position of the KOI
- `koi_fpflag_nt`, `koi_fpflag_ss`, `koi_fpflag_co`, `koi_fpflag_ec`, various flags indicating potential issues with the data gathered for this observation.
- `koi_period`, `koi_time0bk`, `koi_impact`, `koi_duration`, `koi_depth`, `koi_prad`, `koi_teq`, `koi_insol`, `koi_steff`, `koi_slogg`, `koi_srad`, which are all variables measuring properties of the (potential) planet and its star (see the links above for more details). Each of these variables is accompanied by quantities labeled `err1` and `err2`, which quantify uncertainty in the positive and negative direction, respectively. For example, in the first row of the data set, we have
  `koi_insol=93.59   koi_insol_err1=29.45   koi_insol_err2=-16.65`
  which indicates that the value of `koi_insol` is believed to lie in the range $[93.59 - 16.65, 93.59 + 29.45]$.

Assignment Your task is to study one or more of the following questions:

1. Consider the set of variables `koi_period` ... `koi_srad`, measuring properties of the KOI, as described above. Test which variables are conditionally independent of `koi_disposition`, conditional on the other variables in this set. A key challenge here is that there may not be any parametric model that is a good approximation to the joint distribution of all these variables, so you will need to think about other strategies (e.g., a nonparametric approach / a binning method / etc). Be sure to consider and discuss the issue of how approximations (such as replacing continuous variables with binning) could affect validity and power for your approach to this question.
2. Consider the problem of building a classifier mapping from the measured variables `koi_period` ... `koi_srad` (perhaps together with their estimated errors and/or with the KOI position variables `ra` and `dec`, or whatever you choose). The available training data is given by all data points with `koi_disposition` equal to "CONFIRMED" or "FALSE POSITIVE", while the "CANDIDATE" data points are unlabeled (i.e., the test set). However, there are differences between these two portions of the data—for example, we can see that `koi_impact` has a very different distribution among the labeled (training) and unlabeled (test) portions of the data set. Your task is to find a way to control for or adjust for this difference, in order to be able to estimate the accuracy of your classifier on the test set. For this question, it's fine to build a very simple classifier (e.g., based on logistic regression)—the goal of this question is to focus on the challenging inference problem, and we are less interested in the question of finding the best possible classifier.
3. You are also welcome to design your own question(s), but if your group chooses to do this and not any of the options above, then we highly recommend checking in with the instructor/TAs very early on to make sure that your proposed direction works well for the assignment.

Your final report should give a thoughtful discussion of the issues you have identified, and the strategies and methods you developed to address them. You are welcome to use (with citation) any methods that you find in the literature, but you are not required to do this and in general it is more important that your analysis shows a thoughtful and creative approach rather than finding the most effective possible method from the literature.

Remember that the goal of this project is to find questions that are deep, interesting, and challenging *as an inference problem*. For example, developing a creative and effective new method to classify KOIs (i.e., to try to distinguish between true exoplanets & false positives), is interesting and challenging from the point of view of classification/estimation (and from the point of view of the scientific application), but is not a good choice for this project unless it also incorporates some interesting inference questions (i.e., something about quantifying uncertainty for this classifier).

Guidelines Groups of size 2, 3, or 4 are allowed for the mini-project. The extent of the project (e.g., the range of questions explored / methods tried / etc) should be proportional to the size of the group. What you hand in:

- Each group should hand in a written report and either include code throughout the report and/or include the code as an appendix. Please designate a single group member to submit everything on Gradescope, and add the other students in the team group members.
- For your code, it should be clearly organized and commented— for example, you may want to label sections of the code so that we can see which part of the report or which plot/table it corresponds to, add comments to explain steps where notation / variable names / nature of the calculation aren't obvious, etc.
- There are no page length or formatting requirements for the written report. Your report should describe the problems and questions you posed, the details of any methods you implemented / models fitted / hypotheses tested, describe your findings and show plots or numerical results as appropriate, and should discuss some interesting issues relating to inference (for example, multiple testing / appropriately controlling for confounding factors / reducing a high dimensional model to a manageable size / etc). You can also include a discussion of open questions and issues that were not addressed (due to time limitations and/or limitations of the available data).