Machine learning intern -- Take-home exercise

This is intended to be a short exercise (3-4 hr) that allows you to demonstrate your ability to work with data and perform predictive modeling. We'd like to see the way you approach answering business-relevant questions with data. The data used in this exercise is derived from crop model simulations performed by Mandrini et al (2022). We have subset and munged these data to mimic a hypothetical example that closely resembles what you would encounter in real life.

Background and Goals

Field researchers are working on testing new strains of diazotrophic microbes that supply corn with fixed atmospheric Nitrogen (N) to reduce the amount of synthetic N fertilizer needed. For a site to be a good candidate for testing these microbes, however, N availability must be the main factor limiting crop productivity. In previous years, we have seen that about 30% of trials do not respond to N fertilizer additions, meaning that these sites represent a significant drain on our time and resources.

You are tasked with developing a predictive model that helps agronomists assess whether a given experimental site is likely to respond to N fertilizer additions, and therefore be a good candidate location to test our products.

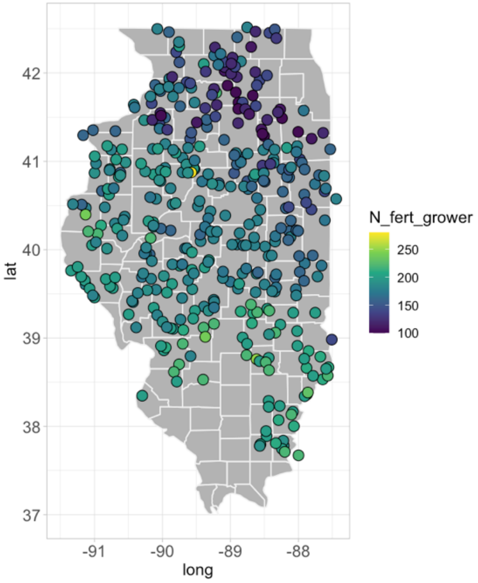
Dataset

The data you are given represents a compilation of 400 field trials conducted in Illinois during 2017 and 2018. At each site, growers were asked to run replicated strip-trials with two treatments:

1) Grower standard practice (GSP, i.e., their normal set of crop management practices)
2) Same as GSP but with a N fertilizer reduction of 40 lb N/a (Reduced N)

Two crop traits were measured at each site:

1) End-of-season grain yield (in bushels per acre)
2) Aboveground plant N at the end of the growing season (in lb N per acre)



The **`targets.csv`** file contains the mean site trait value for each treatment. **Note that prediction targets of interest here are the trait responses (i.e., the difference of GSP minus the Reduced N treatment), not the absolute values.**

The **`predictors.csv`** file contains a set of (potentially) predictive features for each site which we have put together through various data mining efforts. (Descriptions for these predictors can be seen in Table 1 of Mandrini et al (2022)). All predictors are either continuous numeric (floats or integers) or categorical (character strings). Note that about 10% of the data in these predictors are missing (NA).

Task/Questions to Address:

1) Re-format the `target.csv` dataset so that it contains the response of both traits at each site, rather than their raw values. Consider exploring the responses using visualization and descriptive statistics.

2) Load the `predictors.csv` dataset and encode each feature according to their data type. Consider scaling, imputation, dimensionality reduction or other pre-processing you deem necessary before modeling. Be prepared to justify your choices.

3) Note that some of the features are labeled as "junk" to indicate that we know that they have no relationship to our targets. However, let's pretend that you did not know that. Implement a programmatic/algorithmic approach that identifies all potential junk features for removal. Report how many of these junk features your approach was able to identify. Feel free to ignore junk features in the following steps of your analysis.

4) Train supervised ML algorithms **to predict whether the response at a given site will be positive (i.e., greater than zero).** Do this for both target responses (It is ok to fit separate models to each target response). We encourage you to try various model types (e.g., Random Forest, Deep Learning, Regularized regressions, boosted trees, ensembles, etc.) to assess how much the choice of algorithm affects the predictions. We expect you to use reasonable model evaluation procedures based on out-of-bag predictions. Produce a summary table to report model performance for all the trained models.

5) Pick the best-performing model(s) for each target and use it to address the following questions:

   a. Are there key site characteristics (or grouping of site characteristics) that are most strongly associated with whether a site would be responsive to 40 lb/ac of N fertilizer?

   b. Is there any discernible pattern in the learned relationships between the target and the most important features?

   c. If we were to repeat this experiment in the future, what would be the minimum number of sites needed to train a reasonably accurate model?

Outputs:

- A self-contained, repeatable analysis saved in a .zip file. Please script your analysis using either python or R using a jupyter notebook/rmarkdown to generate a static report. Make sure you provide an export of your static report as HTML or PDF. Organize and document the files in this folder such that the logic of data flow and analysis is clear and reproducible.
- Please be prepared to walk us through your report as part of your technical interview (~15-20 minutes). It is likely you may generate more content than can be presented in this time, so consider what might be most important/interesting findings to show to a hypothetical audience of 'stakeholders' consisting of data scientist and field agronomists. Feel free to use a slide deck to summarize your findings if that helps your delivery, though this is optional.

What we're most interested in:

- Clarity of work
- Demonstration of analytical process/reasoning
- Outputs and visuals which might help inform decision making and communication to non-data scientists

Reference:

Mandrini et al. (2022) Simulated dataset of corn response to nitrogen over thousands of fields and multiple years in Illinois. Data in Brief. *https://doi.org/10.1016/j.dib.2021.107753*