Corn Yield Modeling Case Study

Overview

Congratulations on advancing as a candidate for a position within the GeoInnovation Data Science Team. A key part of the job will be developing high quality, predictive models using messy, unfamiliar data and being able to explain to stakeholders what you have done. We have designed a take-home case interview to help us assess (and allow you to showcase!) your skills in:

1. Data preparation and cleansing
2. Exploratory analysis & feature building
3. Construction of predictive models
4. Evaluation of predictive models
5. Presenting your results

Instructions

Download the following 3 files at the links below, let us know if you have any problems with permissions. The files are all gzip compressed CSVs.

- common_data.csv.gz
- daily_observations.csv.gz
- annual_yields.csv.gz

The common data file holds historical yield information for major corn producing US counties from 2003 to 2017. These data can be joined with predictors created from the daily observations file via the “year” and “adm2_code” fields. Hint, the greenup and dormancy dates listed in the common data file might help you to create aggregate predictors from the numerous daily observation values. Additionally, the annual yields file might be employed to adjust the annual county level yields for the national trend of increasing yields.

Your objective is to leverage the data provided to create predictors and a predictive model that can be used to forecast county-level yields for the 2018 growing season. Seeing as the USDA has not yet released 2018 county level yields there is no “truth” to compare against, so breathe easy. There’s no winning or losing based on predictive accuracy, really, we’re most interested in understanding how you approach the problem and if you take steps to build a model based on historical training data that will be predictive for 2018. Cross validation to estimate how accurate you expect your model to be is important.

Building a highly predictive model is only part of the story. We also want to know how you do it. As such, we’ll also be looking for you to deliver a brief (around 10 minute) debrief on your approach to points 1-4 above. Additionally, we’d like you to post your code on GitHub and share a link to it with us.

To recap, we’re looking for:

- Data preparation & modeling code (via GitHub repository)
- Visuals (in format of your choosing) to accompany a discussion on data exploration, modeling process and model results

Best of luck! Please don’t hesitate to reach out with any clarifying questions.

Appendix: Data Dictionary

The following provides a brief description of the contents of each of the data files.

Common Data

- adm2_code: Modified county FIPS code
- adm1_code: Modified state FIPS code
- year: Year of interest
- yield: Yield (in bushels/acre) for year of interest
- area_harvested_obs: acres harvested within the county for year of interest
- phen_gup: Expected greenup date for crops in the county
- phen_dor: Expected dormancy date for crops in the county

Daily Observations

- adm2_code: Modified county FIPS code
- year: Year of interest
- date: Date of observed values
- doy: Day of calendar year, associated with date
- met_avg_t: Observed average temperature
- met_extreme_cold: Boolean flag for extremely cold day
- met_extreme_hot1: Boolean flag for extremely hot day (1)
- met_extreme_hot2: Boolean flag for extremely hot day (2)
- met_gdd: Count of growing degree days
- met_max_rh: Maximum relative humidity
- met_max_t: Maximum temperature
- met_max_vpd: Maximum vapor pressure deficit
- met_min_rh: Minimum relative humidity
- met_min_t: Minimum temperature
- met_p_mm: Observed precipitation
- met_sh: Observed specific humidity
- met_sr_wm2: Observed downwelling shortwave radiation
- mod_evi: Observed Enhanced Vegetation Index
- mod_lst_day: Observed land surface temperature (daytime)
- mod_lst_night: Observed land surface temperature (nighttime)
- mod_nbar_1: Observed MODIS band 1 reflectance
- mod_nbar_2: Observed MODIS band 2 reflectance
- mod_nbar_3: Observed MODIS band 3 reflectance
- mod_nbar_4: Observed MODIS band 4 reflectance
- mod_nbar_5: Observed MODIS band 5 reflectance
- mod_nbar_6: Observed MODIS band 6 reflectance
- mod_nbar_7: Observed MODIS band 7 reflectance
- mod_ndvi: Observed Normalized Difference Vegetation Index

Annual Yields

- year: year of interest
- yield: national yield for year of interest