Agritask

Hi,

In this home test you have been given two datasets containing information about soybeans crop yield (actual yield for each field in kg per hectare), weather data, soil data (type of soil and soil measurements) and remote sensing data (various vegetation indices taken from satellite images) for soybeans crops in Brazil. Your task is to build a model to predict soybeans crop yield for a given set of available data.

Data

There are two data sets attached (both are **tab separated csv files**) –

- seasonal_dataset – data which represents a full season (from planting to harvest)
- seasonal_phen_dataset (phenological data) – data which represents a full season (from planting to harvest), and aggregated by the phenological stages of this crop (Soybeans) - same data as the 1st dataset, just different feature engineering.

Columns –

- plot code – ID of the plot (field) for which yield and measurements are stored
- Season code – the season for which the data is stored (each plot could have few seasons)
- "Sitlav - Soybeans - Yield - KG/Ha" – actual yield that was harvested in that plot, in that season (in kg per hectare) – **this is our Y which you are tasked to predict**.
- Columns of spectral data which are measured over a season of ~ 80 days and aggregated for a full season using moments (i.e Savi2, NDWI etc.)
- Soil Type - class ("CLASSE_DOM")
- soil measurements data - columns such as "**mean_cec_5_15", "mean_clay_0_5", "mean_sand_15_30"** – measured through the season.
- seasonal temp/rain/humidity - which are measured over a season of ~ 80 days and aggregated for a full season using moments (i.e Rain, Daily Mean temp, etc.)

phenological data –

- almost same data as the seasonal, just aggregated by important stages of the plant life cycle (phenological stages)
    - phen1 - Crop Establishment - 0 to 20 days
    - phen2 - Vegetative Growth - 21 to 55 days
    - phen3 - Early Reproductive Stages - 56 to 70 days
    - phen4 - Grain Filling Stages - 71 to 81 days
- In this dataset, there are additional agro-climatology parameters which don't appear in the 1st dataset - total dry days, max dry days, total rain, gdd - per phen stage.

Requirements

1. Conduct exploratory data analysis to understand the data and its distribution. Include at least two visualizations that you find helpful.
2. Feature Engineering: We have already done some feature engineering so this could be skipped. But feel free to create any additional features that you think would be useful for predicting crop yield. **Also think about what you could have done differently here**.
3. Model Selection: Build at least two models to predict crop yield. You may use any libraries or packages that you prefer.
4. Model Evaluation: Evaluate the performance of the models using appropriate metrics. Include visualizations to help communicate your results. This should be done for both datasets, then show what dataset works best (i.e. what is the better feature engineering)
5. Model Tuning: Fine-tune the best performing model and provide a justification for your choices.
6. Business Insights: Provide recommendations on how the model can be used to optimize crop yield, including any limitations or potential biases that should be considered.

Good luck! And please let me know if you have any questions.

Cheers,
Kiril