Predict Yield for each farm ID

Approach

Goal:

As a part of this challenge, our task at hand is to build a high performance and interpretable machine learning model to predict the yield based on the farm data and weather data.

About data: We are provided with the sample datasets of the different farms holding the information of farming company and location of that farm, farm area and number of processing plants that farm have and ingredient type that each farm producing the yield. The train data consists of 20216100 rows and 4 columns, and the test data consists of 20848800 rows and 4 columns. Actually it has time variable the yield is recorded for every day every hour in one year we need to forecast for next year. We can solve this problem in regression also. The target variable is a continuous variable hence, making the problem a Regression Problem.

Below given are the list of features and what it represents:

1. Date - every hour data for every day in year 2016
2. Farm id - unique farm ids representing each farm
3. Ingredient type – type of ingredient that each farm producing the yield
4. Yield - yield of each ingredient type in one year (Target variable)
5. Deidentified location – location of each farm
6. Temp_obs - Temperature observation at each location
7. Cloudiness – amount of cloud cover in the sky at a particular time and location. Clouds can affect the amount of sunlight and temperature received by crops, which influence the photosynthesis and growth
8. Wind direction – observation of wind direction in each location
9. Dew temp – The temperature to which air must be cooled (it reduces the heat stress in very hot environments)
10. Pressure Sea level – Atmospheric pressure at sea level at a given location. If it is high pressure, we may get dry weather if it is low we may have wet weather
11. Precipitation – form of water (rain, snow) that falls to the earth's surface
12. Wind speed – speed of wind, strong winds can damage crops by breaking stems while gentle air can help to improve air circulation and reduce the diseases.
13. Operations commencing year – farm first began operating. The age of farm can impact the quality of the soil and the infrastructure such as irrigation systems in that place. Older farms may have outdated or insufficient equipment while newer facilities may have access to more advanced technology and equipment. It can also influence the experience and knowledge of the operators

14. Number of processing plants – total number of facilities that are involved in processing a particular ingredient. It includes cleaning, drying, milling, or refining. more processing plants indicate that higher demand for that crop, which can lead to increased production and higher yields.
15. Farm Area – total area of land that is used for agricultural production. Higher farm area indicates a greater capacity for production
16. Farming company – company that involved in agricultural production. Farming company with experienced managers and access to advanced technology may be able to achieve higher yields than a company with less experience and fewer resources.

Exploratory Data Analysis: The first thing in EDA is to check whether there are any null values present in the data. In the data provided there were null values in the weather data and farm data. Then after, I separated the categorical and numerical features to perform analysis

The data contains majority of numerical features. The following is the list of categorical features and its unique values:

1. The farm id contains 1434 number of unique values
2. The ingredient type contains 4 unique categories
3. The farming company contains 16 number of unique categories
4. The location contains 16 number of unique categories

The numerical feature was further separated into discrete and continuous features. Graphs showing the relationship between the numerical features and the target variable were plotted. There is some number of outliers present in the data, which contributes positively to the data. So, I kept it as it is in order to prevent the loss of information.

Data Pre-processing / Feature Engineering: The data pre-processing step consists of the one-hotencoding of the categorical features to convert it into numerical features and scaling/normalizing of the data so that all features are in the same scale. I divided the training data into training and validation data. here I have used both one hot and label encoder to convert categorical to numerical variables and used min max scaling to get the data in same scale. Finally, the data after the preprocessing step, is ready to be fed into the models.

Model Creation & Hyperparameter Tuning: Various Regression models were created and hyperparameter tuning was done in this step.

The models used were: 1. Linear Regressor 2. Random Forest Regressor 3. XGBoost Regressor 4. Gradient Boost Regressor 5. Light gradient boosting regressor.

The models were trained on the training data and were validated on the validation data just to see which models are performing well on unknown data. It was observed that ensemble models were performing well. Finally, the model performing the best on validation set was selected as the final model.

Finalizing the model: The final model was trained on the original test data set and the predictions are made. Final predictions are stored as a csv file for the submission.

Here I have done hyperparameter tuning using randomized search cv but unfortunately, with the available resources I am unable run it on my dataset.

To increase the yield for ingredient_w I will take following measures:

1. I will conduct a **root cause analysis.** I will investigate why ing_w is not giving any yield. This may involve analyzing the production process, reviewing quality control measures, and identifying any issues with the raw material or supplier.
2. I will **improve quality control measures**. I will establish and implement quality control measures to ensure that ing_w meets the required quality standards. This may include testing the raw materials before processing, monitoring the production process, and inspecting the finished product before packaging.
3. I will **optimize processing conditions**. Review the processing conditions for ing_w to identify opportunities for optimization. This may involve adjusting temperature, pressure or processing time, or modifying the equipment or machinery used.
4. I will consider **alternative suppliers or ingredients:** if the quality control and optimization do not lead to improved yield, consider alternative suppliers or alternative ingredients that may be more suitable for the product and the production process
5. I will **monitor and review the performance**. Regularly monitor the performance of ing_w and review the results to identify areas for improvement and ensure that the steps taken are effective in increasing yield.