Supporting Document

1. **Describe the key elements of your approach and assumptions of your approach if any; and What is innovative and unique about our model ?**

Our methodology involved extracting specific spectral bands from satellites, including Sentinel 1, Sentinel 2, Landsat 8, and Landsat 9. These bands were used to compute vegetation indices that capture critical growth features of rice, and an index on ground soil moisture, which provides insights into the water content of fields earmarked for harvesting.

Additionally, to account for external factors such as weather, we incorporated a supplementary dataset containing weather data specific to the regions of Chau Phu, Chau Thanh, and Thoai Son. This dataset enabled us to gain access to precipitation patterns in the areas of interest, which is a crucial factor in predicting rice yield. It is worth noting that all crops require water during their germination stages to achieve optimal growth and yield production, with precipitation being the primary source of water for crops.

Our approach utilized a distinctive time window that matches the growth cycle of rice leading up to harvest. Nevertheless, our model can be flexibly applied to estimate rice yield on any given date, as our time window accommodates the satellite's orbit on each specific day of interest and thus gives reliable results. This approach helps us to capture the subtle details of the rice plants and accurately reflect the actual conditions on the ground. Essentially, we are able to see visible patterns in a narrow time frame. Our intuition here is solely based on the fact that we are addressing the issue of food security; as we are considering developing country contexts, where there are limited bandwidths and resources available to analyze larger amounts of data, our aim has been to reduce memory usage and need of processing power, while still maintaining accurate results.

Our methodology is designed to minimize the level of noise in vegetation index calculations by eliminating the reliance on larger time windows or entire crop cycles. This approach reduces the potential for extraneous data to be incorporated into the calculations, thereby enhancing the precision and accuracy of our index calculations for a specific date. By using this approach, our calculations are better equipped to capture the nuances of the target vegetation and reflect the actual conditions on the ground around the time of observation, resulting in more reliable data for analysis and decision-making purposes.

To improve the efficiency of our model, we used multiple regression models and aggregated their outputs to generate a final prediction. We employed three ensembling approaches, including Bagging, Boosting, and Stacking, to improve accuracy and robustness. Bagging was selected based on the highest evaluation score on the leaderboards, and it consistently produced accurate predictions when compared to other popular ensembling techniques like boosting and stacking. Our bagging approach utilized Random Forest, XGBoost, Gradient Boosting, and CatBoost, which provided diversity, strong performance, robustness, and complementary strengths. By combining these models, we achieved an ensemble that benefited from their combined strengths, such as the ability of XGBoost to handle fast calculations, CatBoost's capacity to manage categorical features, Random Forest's ability to manage noisy data, and boosting algorithms' ability to capture complex relationships.

A few of the assumptions we made for our model were based on taking the mean of all the vegetation indices.For each date, we received multiple values and we concluded that the data on the satellite is based hourly and not daily. So we took the mean assuming that on that particular day the vegetation indices and ground soil moisture values did not shift by a lot.

We used cubic spline interpolation to replace the missing values and since we did not know the actual values to refer to we assumed that our interpolation was correct.

While calculating some of the vegetation indices, some of the coefficients vary according to soil type and it could be further enhanced by knowing the topology of Vietnam, however finding such relevant data was very difficult and time consuming, so we took an assumption of using a general estimate of all these coefficients by referring to similar soil types found in past research papers.

2. What is our Target and Predictor Dataset?

The target dataset used for our model were the Rice Yield column values from the Crop yield dataset to train and test our model for prediction.The Date of Harvest column values from the same dataset were utilized to find precipitation data from our additional dataset and acted as filters for reducing the data volume from our satellites using our unique small time window feature. Latitudes and Longitude columns from Crop yield were used to extract and structure our predictor variables.

For the predictor dataset, the spatial band values that we utilized to calculate vegetation indices and ground soil moisture, derived from Sentinel 1, Sentinel 2 and Landsat 8,9 were used.

We chose NDWI, NDVI, SAVI, EVI2 and Albedo to calculate from Landsat 8,9 spatial bands with a time frame of 16 days (8 days before and after Date of Harvest)

- NDWI (Normalized Difference Water Index) = (NIR - SWIR) / (NIR + SWIR), with NIR representing the near-infrared band and SWIR the shortwave-infrared band.
- NDVI (Normalized Difference Vegetation Index) = (NIR-Red) / (NIR+Red).
- SAVI (Soil Adjusted Vegetation Index) = ((NIR - Red) / (NIR + Red + L)) * (1 + L), with L being a soil adjustment factor (typically between 0 and 1).
- EVI 2 (Enhanced Vegetation Index 2) = 2.5 * ((NIR - Red) / (NIR + 2.4 * Red + 1)).
- Albedo = 0.356 * Blue + 0.130 * Green + 0.373 * Red + 0.085 * NIR + 0.072 * SWIR +0.0018

From Sentinel 2 spatial bands, we chose to calculate FAPAR and LAI with a timeframe of 20 days (10 days before and after Date of Harvest)

- FAPAR (Fraction of Absorbed Photosynthetically Active Radiation) = (1 - (NIR - Red)/(NIR + Red)) * (1 - 0.98 ** (1 / (0.0038 * Red))).
- LAI (Leaf Area Index) = 0.618 * ((NIR - Red)/(NIR + Red) * * 1.334)

Finally from Sentinel 1 spatial bands, we calculated soil moisture as the only feature with a time window of 12 days (6 days before and after the Date of Harvest)

- Soil Moisture = (1 - ((10 ** (0.1 * DOP)) / A) ** B) where DOP is degree of Polarization that is found as DOP = (VV / (VV + VH))

From our Additional Dataset from Visual Crossing, we received the precipitation values directly, hence no calculation was required and a time window from 1st October 2021 to 1st September 2022 was used as it covered the growth cycles at least 100 days before each Date of Harvest within this time window.

3. How did we treat missing values from Satellites ?

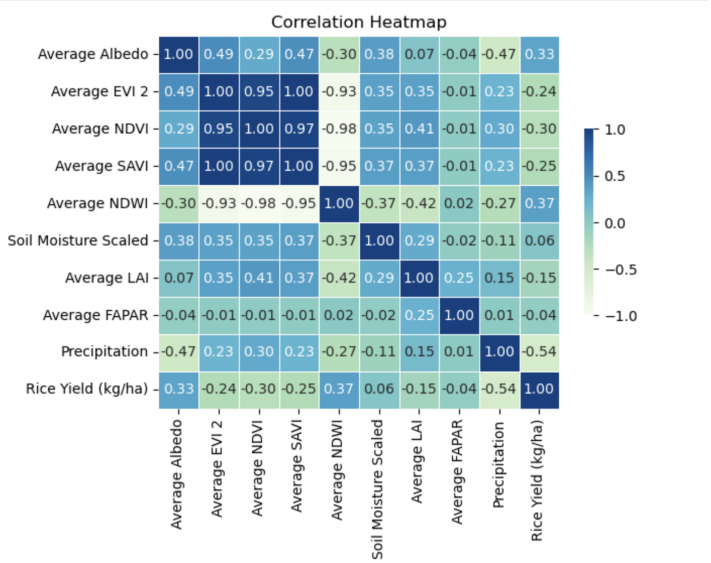
To replace NaN values in our satellite data we used the Cubic Spline Interpolation. Cubic spline interpolation is a method of interpolating data points with a smooth curve, using a piecewise cubic polynomial function. It can be used to replace NaN values in a dataset by interpolating between the adjacent valid data points. This method is better than other methods of replacing NaN values as it generates a smooth curve and is less sensitive to outliers. Replacing the missing values with zero, mean or any other statistical feature was introducing greater bias in our predictor variables.

4. How did we generally treat outliers?

In our case, scaling the input features was deemed non-essential, as we were primarily calculating indices between 0 and 1, or between a strict range of values. However, as an improvement we would still want to standardize input features to promote cohesion and mitigate the impact of (if any) outliers.

5. How was correlation checked?

In order to see how our predictor variables were correlated with each other, and with our target variable we plotted a correlation heatmap which is shown below:



6. How were features selected and extracted for the ML algorithms used ?

The features for each ML algorithm used were selected using this correlation heatmap, feature importance plots and trial-and-error. EVI2 and SAVI are highly correlated with each other, and NDVI. This was introducing multicollinearity in our model. Additionally, EVI 2 and SAVI had the lowest feature importance and so, we removed those features from our data to prevent overfitting of our ML model. These final sets of features were overfitting-proofed, and were accurately explaining crop growth attributes. On our way to selecting these features, we came across many methods such as Recursive

Feature Estimation and Principal Component Analysis. However, these methods were either not reliable or could not be generalized for the entire model at hand.

7. **How did we divide our data for training our model and what did we do for model validation?**

We divided our Target dataset "Crop_Yield_Data.csv" into train and test data in a 80/20 ratio and the challenge submission template "Challenge_2_submission_template.csv" assigned to each team was later used as the validation data for our model. Since the dataset is smaller, and we had unseen data to test the model, we chose an 80/20 split instead of 75/25. We ran our model through the Target dataset with the predictor dataset to train our model. Then we used the test data to evaluate performance. Upon reaching a good enough R^2 of greater than 0.50, we proceeded with the validation data and made our initial submission to the EY leaderboards. We then went on fine tuning our parameters for each model and ensembling to get better results.

8. **How was the performance of the model measured and shown?**

The performance of our model was measured and seen using R^2, mean squared error, feature importance plots, scatter plots and residual histograms. For our python notebook, we have showed the R^2, mean squared error and feature importance plot (for bagging).

9. **What regression techniques were used to make the model?**

The ML algorithms used for predicting rice yields were Random Forest Regression,Gradient Boosting Regression, XGBoost Regression and CatBoost.

- Random Forest: Random Forest is a powerful algorithm that is resistant to overfitting and can handle a large number of input features. It can handle non-linear relationships, high-dimensional data, and automatically perform feature selection.

- Gradient Boosting: Gradient Boosting is a potent algorithm capable of handling both regression and classification problems, achieving remarkable accuracy across diverse datasets. It excels at capturing complex relationships and can accommodate a wide range of data types.

- XGBoost (Extreme Gradient Boosting): An optimized version of Gradient Boosting, XGBoost is more efficient, scalable, and offers regularization options to help prevent overfitting.

- CatBoost: A gradient boosting library that specializes in handling categorical features, providing strong performance while reducing the risk of overfitting. It's fast, efficient, and can handle a mix of categorical and continuous features.

In the end, we bagged these ML models. Bagging, short for Bootstrap Aggregating, is an ensemble learning technique used to improve the performance and stability of machine learning models. It combines the predictions of multiple base models to produce a more accurate and robust aggregated prediction. Bagging is particularly effective in reducing the variance and overfitting of models that have high sensitivity to small changes in the training data, such as decision trees.

10. How was model selection done and what was the model's performance?

To evaluate bagging, we compared it with boosting and stacking algorithms. Boosting and Stacking out-performed bagging in the training phase, but were not accurate enough on unseen data. On the other, bagging was consistent. After extensive trials and cross-validation, our best leaderboard score was achieved using four base models: Random Forest, XGBoost, Gradient Boosting, and CatBoost.

Results of our individual ML model:

```
XGBoost Regression Accuracy: 0.6084
XGBoost Regression MSE: 460.1405

Random Forest Regression Accuracy: 0.5842
Random Forest Regression MSE: 474.1525

Gradient Boosting Regression Accuracy: 0.6063
Gradient Boosting Regression MSE: 461.3786

CatBoost Regression Accuracy: 0.6395
CatBoost Regression MSE: 441.4552
```

Results of Bagging technique:

```
Boosting Regression Accuracy: 0.6323
Boosting Regression MSE: 445.8798
```

11. Describe your highest performing features

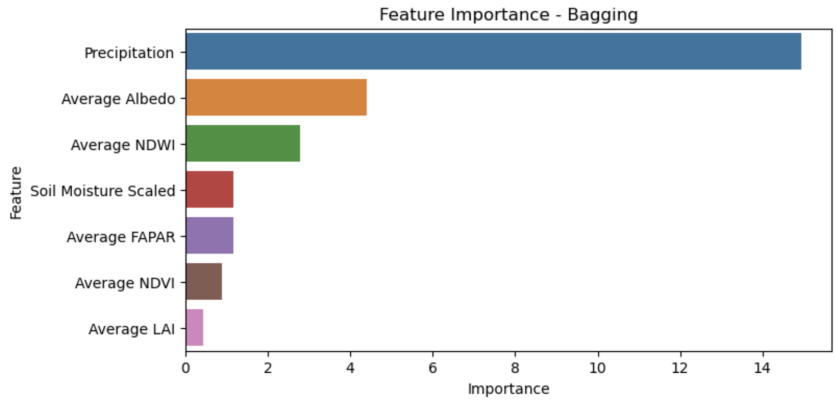


Figure : Importance plot for predictor features

We see that Precipitation, Albedo and NDWI are the highest performing features. Focusing on optimizing these factors may lead to improved crop yields and contribute to enhanced food security in the region. Here's how:

1. Precipitation: Adequate rainfall is essential for rice cultivation, as it provides the necessary water for crop growth. By understanding the relationship between precipitation and rice yield, farmers can make informed decisions about irrigation practices and water management, ensuring that crops receive the right amount of water at the right time.

2. Albedo: As mentioned earlier, Albedo measures the reflectivity of the Earth's surface, which affects how much sunlight is absorbed or reflected by the crop. By optimizing the crop's Albedo,

farmers can ensure that rice plants absorb an adequate amount of sunlight for photosynthesis, leading to more vigorous growth and higher yields.

3. NDWI: It is an indicator of water content in vegetation. By monitoring NDWI, farmers can assess the health of their rice crops in relation to water availability. This information can be used to make better-informed decisions about irrigation and water management, helping to maintain optimal water content in the plants and potentially increasing yields.

By developing targeted agricultural strategies and resource allocation plans based on these influential features, we can help farmers and agricultural organizations in Vietnam optimize the conditions necessary for successful rice cultivation. This, in turn, can lead to improved crop yields and enhanced food security for the region.

12. Describe the Evolution of your Approach for selecting the Best Model

To realize the best features for each of our regression models we initially performed Cross Tabulation. This allowed us to see the correlation coefficients of our predictor variables with respect to rice yield but upon further brainstorming, we saw that the Correlation Heatmap was a better method that does exactly what Cross tabulation does and shows it in a much more visualizable way.

Principal Component Analysis(PCA) was also done to try to reduce the dimensionality of the data but there was no way to figure out which predictor variables fall onto the individual principal components and since we were focused on predicting rather than categorizing rice yield, PCA was not a suitable way to reduce dimensionality in our case. Instead, we figured out how to filter data from our satellite query by using our unique time window, which ultimately made use of PCA obsolete.

Furthermore, we performed the Recursive Feature Estimation method for our predictor features. However, using the Recursive Feature estimation, we found that some regression techniques we used did not have similar features and performing individual feature selections meant that the Bagging of all these ML algorithms could not be smooth out and generalizability of our model would be reduced.

Therefore, for feature selection we ended up using the correlation heatmap and feature importance plots, and for extraction we used our unique small time window. And lastly, we simply adopted a trial-and-error methodology after analyzing our evaluation scores. We still had a low evaluation score on the EY leaderboards and so we decided to look into the values of our individual features themselves and realized that there are a lot of missing values from some satellites and some values were not in their valid range.

We brainstormed a little and figured out a way to improve our features further by replacing the NaN values using cubic spline interpolation. We scaled feature values that were out of their valid range into their desired ranges using the MinMaxScaler function and tried uploading submissions again.

Initially Precipitation was not included into our predictor variables and so we looked into specific weather attributes that may affect rice yield directly and we found out that amount of precipitation in the form of rain directly influences rice yield and we verified this with multiple research papers. So we ended up finding a website namely Visual Crossing that allowed us to get precipitation on a daily basis and

according to our preferred timeline. We also looked into MODIS which is NASA's weather dataset however, finding data on a daily basis and on the specified region was time consuming and very difficult to look for. This significantly increased our evaluation score up to 0.62and so we ended up retaining this.

From Sentinel 2, we added FAPAR and LAI indices after we received a score of 0.62 on the leaderboards and the reason we held onto Sentinel 2 was the fact that our initial understanding of using all satellites to calculate vegetation indices was not accurate. For example, we can calculate NDVI from both Landsat and Sentinel 2 but Landsat had a higher revisit time which allowed us to get more data around the date of harvest compared to that of Sentinel 2. Sentinel 2 would give us better resolution and so indices like FAPAR and LAI that would benefit from higher resolution should be calculated from Sentinel 2 and not Landsat. So we implemented this only after exhausting all our options to improve our evaluation score.

Later on, we tried hyper parameter tuning the number of estimators, depth, verbose and learning rates of our individual ML models and did trial and error until our R^2 values significantly improved. We also had a backup approach of using Neural Networks more specifically Deep Neural Networks and we tried to predict rice yield using that but it had worse outcomes and negative R^2 than all our ML models combined, and hence we backed away from that approach to save time.

13. What was the Prediction Accuracy Score on the Unseen Data

Our final results for the prediction accuracy on the unseen data on which the model was validated was 0.65 in the leaderboards ranking. We also realized that no matter how big our R^2 was, it did not show any positive correlation with the evaluation score received on the EY Leaderboards. So we kept on brainstorming on improving our model with less reliance on the R^2 values itself and more on the evaluation scores received.

14. What was our most important breakthrough?

The most important breakthrough that helped us a lot to improve our score was to use the additional dataset's precipitation values. Our team had previously seen precipitation as a recurring feature in many past research papers, however there was no direct way of calculating precipitation from the available satellites directly. Many of the past papers talked about using MODIS however, finding precipitation data according to the day was very difficult and since there was a large volume of data to choose from, we ended up going for a simpler weather data acquisition website Visual Crossing. We realized the importance of precipitation when we added this as a feature and trained our model. Our evaluation score jumped to 0.62 from here and precipitation is also shown to be the highest performing feature out of all our features.

15. What was the hardest thing about solving the problem?

The hardest thing about solving this problem was to find appropriate features for our ML models to begin with and how to address those missing values in the dataset. It took a lot of time,research and dedication to fill the gaps in our knowledge for predicting rice yield. Many of the features that could have greater potential in predicting rice yields such as ground truth rice crop lengths cannot be directly calculated using our satellite's spatial bands and hence could not be considered.

Reducing the amount of data we were getting was also difficult and hence we went for the smaller time window around the date of harvest. Sometimes, some satellites especially Landsat would give an API error as the satellite was down and so it killed a lot of our time that we could have otherwise utilized on improving our model.