# "Solar Radiation on Agricultural Soil Conditions Using Data Science and AI Models"

Author: Zonghong Yu*

*DSAN5550 ** Georgetown University: Data Science & Analytics *04/25/2024

https://github.com/zy236yuz5/DSAN-5550-PROJECT-zy236

## Problem Statement

Agriculture plays a great part in our daily lives. It provides us with valuable food. However, the climate and many other factors affect the agricultural conditions. For instance, solar radiation plays a crucial role in determining soil temperature and moisture levels, which in turn significantly affects agricultural productivity. In West Palm Beach, the solar radiation is very high compared to other places, this also has huge impacts on the soil which in turns affects the crop yields. Therefore, in order to have a better idea about the relationship between solar power and soil condition in different places, this project utilize data science and machine learning techniques to investigate the impacts of varying levels of solar radiation on soil temperature and moisture as well as predictions for soil condition and crop production. By understanding these impacts, we can offer insights into optimizing agricultural practices for better crop yields and more sustainable farming methods in the context of changing climate conditions.

## Data & Methodology

- **Dataset Gathered:**

Data is sourced from VisualCrossing with the weather query builder. Different dataset is obtained for 10 distinct locations: Boston, New York, Atlanta, West Palm Beach, Chicago, Los Angeles, Seattle, Roswell, and Lakewood. The key features are solar energy, solar radiation, soil moisture, and soil temperature. For solar radiation, we also detailed into Direct normal radiation, Diffuse normal radiation, Global horizontal radiation, and Global tilt radiation. In addition, another dataset called Export Price Index (NAICS): Crop Production is also obtained from FRED for making predictions.

- **Dataset Cleaned:**

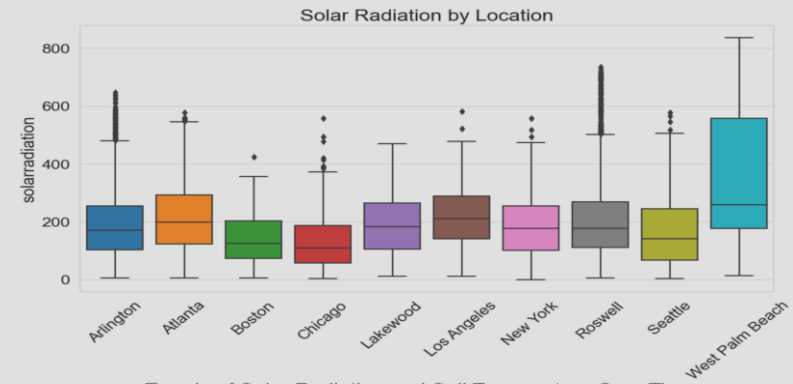
Dataset columns are imbalance due to the different conditions in different places. Therefore, we remove columns that have substantial number of missing values. Then, we remove rows with missing values in the soil and solar-related columns to make the dataset clean, balance, and easy to understand the features. Missing values are dropped due to vast time range, it will not be feasible to replace them.

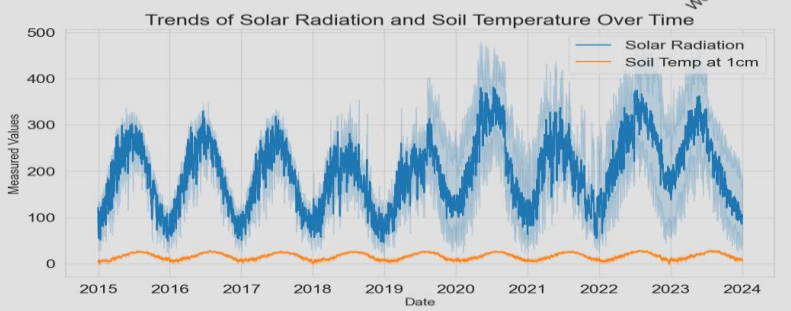
- **Methods & Models:**

By utilizing different models, we can uncover patterns and relationships within the data and make predictions about various outcomes, and ultimately gain a deeper understanding. The methods include Logistic Regression, Random Forest, and Gradient Boosting for Regression or Classification. In addition, we use deep learning methods: LSTM, RNN, and GRU for prediction. The evaluation matrices contain, RMSE, MSE, Confusion Matrix, Feature Importance, and Carbon Cost.

## Exploratory Data Analysis

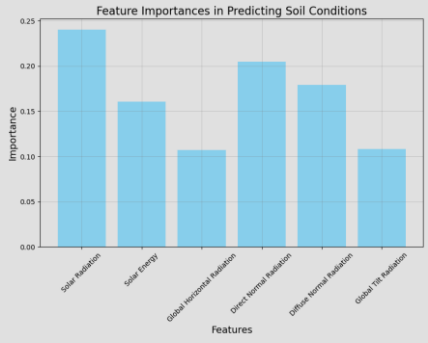
Here, we have the solar radiation based on different locations as well as the trends of solar radiation and soil condition over time. Locations like Arlington, Seattle, Roswell and West Palm Beach have occasional days with extremely high solar radiation compared to others. And It is obvious that they have similar trend.

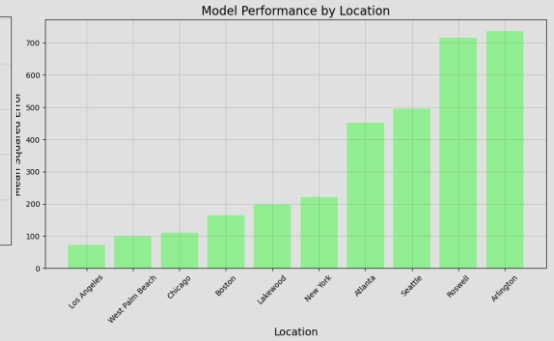


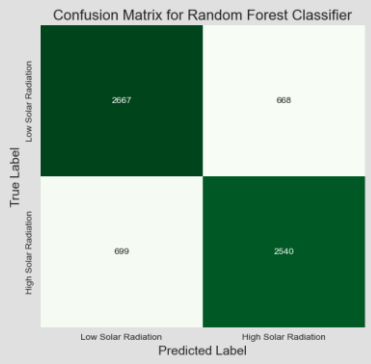

## Evaluation Metrics and Results

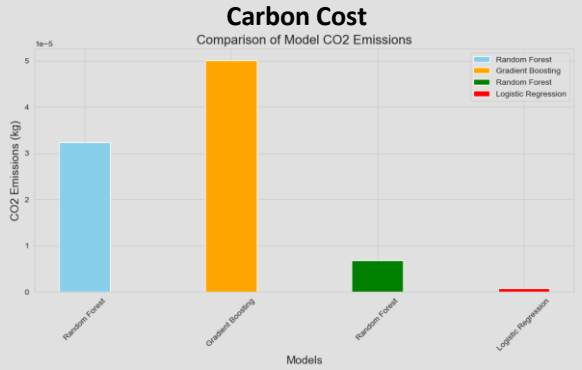
Based on the Random Forest and Gradient Boost, we have found that solar radiation and direct normal radiation are more important in predicting the soil conditions. In addition, we also found that the model performed the best for Los Angeles with the smallest MSE. On the contrary, model performance for Arlington and Roswell results in higher MSE.




Then, we also have other results based on the Random Forest Classifier and Logistic Regression. The accuracy is higher for Random Forest method. However, the carbon cost for Random Forest method is higher than Logistic Regression. The Gradient Boosting method has the highest Carbon emissions.




**Deep Learning Methods Comparison**

The results show that GRU model performs the best in making predictions regarding the crop production. However, it also has the highest carbon cost. As for LSTM and RNN, RNN performs better with smaller RMSE and smaller Carbon cost.

| | RMSE | MSE | Carbon Cost(kg) |
|---|---|---|---|
| LSTM | 0.1087 | 0.01182 | 2.87e-05 |
| RNN | 0.09561 | 0.00914 | 2.08e-05 |
| GRU | 0.00645 | 0.0004 | 3.12e-05 |

## Conclusions, Limitations, and Future work

- We have found that the key features that affect soil conditions are solar radiation and direct normal radiation. And it is more effective for us to make prediction on soil condition in Los Angeles.
- We find that both regression and deep learning method are giving us a great result in term of predicting our target with relatively small carbon cost.
- The predictive ability of the models indicates a strong relationship between the soil condition and solar radiation levels. This relationship can be utilized for predictive maintenance in agricultural activities, optimizing conditions for crop growth, and studying environmental impacts.
- However, there are certainly other external factors such as wind energy, humidity, cloud cover, and so on that can have impact on the soil condition. And crop production is also related to other factors concerning the whole ecosystem. Therefore, more work can be done in the future.
- For future work, we plan to add more features mentioned above to improve the analysis with more datasets that will make the model and result more effective. In addition, we plan to reach beyond the soil condition to include plant-specific data, which can help in creating a more integrated model of environmental monitoring.

## Reference

Visual Crossing. (n.d.). Weather Data Services. Retrieved February 21, 2024, from https://www.visualcrossing.com/weather/weather-data-services

Federal Reserve Bank of St. Louis. (n.d.). Export Price Index (NAICS): Crop production [IY111]. Retrieved from https://fred.stlouisfed.org/series/IY111