A Proposal for Exploring Optimal Machine Learning Models for Predicting Crop Yield at Township Level

By
Navdeep Singh Saini, Mayukha Bheemavarapu, Mehul Bhargava, Val Veeramani

Master of Data Science Capstone

University of British Columbia Okanagan Campus
May 5th, 2022

Background:

Agriculture and Agri-Food Canada is a department of the government of Canada that focuses on the regulatory, production, marketing, and policy aspects of all farms and Agri-based goods. AAFC works to grow Canada's exports, while providing leadership in the expansion and development of a competitive, innovative, and sustainable Canadian agriculture and agri-food sector. Scientific research which covers a wide range of areas including agronomic, environmental, and economic interactions is carried out from more than 20 research and development and virtual centers spread out across the country.

Motivation and Purpose:

Crop yield and production are crucial to meet the demands of millions in this country. Apart from that, one of the most important aspects is to increase exports, and to achieve that, prior information regarding the yield is highly significant, especially for the decision makers who can further formulate the export policies based on the results generated related to the crop yield forecast. Due to crops being one of the 4 key sectors managed by Agri-Food Canada, it is in our best interest to make sure that this sector is performing to its highest capacity. Not to mention that there are several millions of individuals counting on its sustained performance, we are tasked with ensuring that we're able to keep up with the demand of them in the coming years and the future.

As AAFC works to grow exports, while providing leadership in the expansion and development of a competitive, innovative, and sustainable Canadian agriculture and agri-food sector, the project is quite essential to come up with data driven planning with regards to agricultural production.

Literature Review:

Luca Sartore et al. (2022) proposed a methodology that provides a set of artificial covariates created by extracting most of the information from the empirical density functions of real-life phenomena estimated at the county level. This allows relevant features of empirical densities to be used as input variables in standard machine learning algorithms. This approach has been shown to be capable of generating artificial covariates that can be effective when used in predictive models for crop yield predictions at the county level. However, when enriching the model by including higher moments of the distributions inferred at the county level, the proposed methodology cannot downscale crop yield predictions to finer spatial resolutions (such as at the pixel level). Depending on the availability of satellite imagery, finer resolution remotely sensed data can be linked to precision agriculture data for more detailed analyses.

Huiren Tian et al. (2021) developed a LSTM deep neural network that takes advantage of multi-feature inputs that are based on remote sensing data and meteorological data and multi-time steps to improve the accuracy of estimating yield at the county level in the Guanzhong Plain. They used two types of remote sensing data: Vegetation Temperature Condition Index (VTCI) and Leaf Area Index (LAI) and developed a LSTM neural network model for estimating wheat yield. They compared their work with other models (BPNN and SVM), and evaluated the applicability, robustness, and effectiveness of their proposed model. They have also concluded that the LSTM model has better adaptability to interannual fluctuations in the climate.

Umer Saeed et al. (2017) proposed a study aimed to combine weather data with remotely sensed vegetation indices for yield forecasting, using a non-parametric, empirical model based on Random Forests (RFs). This study was designed with the objective to integrate data of minimum and maximum temperature, rainfall, and sunshine hours with MODIS-derived NDVI in different ways for the Punjab province of Pakistan to forecast wheat yield three weeks before harvest using RFs. The second objective of the study was to improve the performance of existing yield forecast models, which are based on NDVI only.

Aston Chipanshi et al. (2015) assessed the skill and the reliability of the Integrated Canadian Crop Yield Forecaster (ICCYF), a regional crop yield forecasting tool, at different temporal (1–3 months before harvest) and spatial (i.e., census agricultural region – CAR, provincial and national) scales across Canada. They found that the forecast reliability improved over the cropping season when more near-real-time data became available. A skillful forecast using the ICCYF could be expected around mid-August. They also attempted to integrate several sources of information during the growing season for forecasting agricultural yields at different administrative levels of aggregation and discussed the role of their methods to complement early-season survey estimates.

Aims and Objectives:

The project aims to explore optimal machine learning models for predicting crop yield based on growing patterns at the township level.

The project will seek to test machine learning (ML) methods for predicting crop yield at the township level. Currently, crop yield prediction and forecasting are accomplished using the Canadian Crop Yield Forecaster where weekly low resolution (~1km) NDVI (Normalized Difference Vegetation Index) values and daily agrometeorological variables from unevenly distributed climate stations are used in a blended model which uses statistical and biophysical modules. Several issues arise when statistical based models are used to predict crop yield, such as multicollinearity among predictor variables and a lack of plausible scientific explanation for some predictor variables that are selected in some experiments (inference). Studies with ML algorithms have shown that predictions can be made without making any prior assumptions about the relationship between the response and predictor variables. Using the Canadian Prairie domain as the study area, we seek to test the following algorithms (or better ones) in the prediction of canola, spring wheat, and barley. XGBoost and Lasso and compare the results with the observed values.

The project will use the Python programming language with its various libraries and packages for applying Deep Learning methods and algorithms.

Tools:

We propose to use Jupyter Notebook/Lab, Google Colab, or any other Python interpreter and implement complicated procedures such as Deep Learning libraries for training the models. We will be using GPUs for faster processing of large-scale datasets.

Desired Data Product/Outcome:

1. We anticipate that crop yield prediction models from ML algorithms that can address agroclimatic conditions as they evolve during the growing season will be tested and recommended for the study area.
2. We will present the model testing results with some well acceptable model performance indicators, such as mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and the coefficient of correlation (R^2) with the training and testing datasets by comparing the model generated yields with the crop yield provided to us.
3. We expect that this study will provide a platform for quickly testing and interpreting new earth observation data sets that will provide physical meaning to outputs with little or no supervision.
4. This study will enhance our understanding of the spatial and temporal variability of crop yields for early warning purposes.
5. Supporting program and policy development as Canada addresses climate extremes and change.

Dataset:

The project involves predicting the crop yield at a township level for the provinces Manitoba and Saskatchewan. The total agricultural area is divided into individual entities of 10x10 sq.km and each of such individual unit is called township.

To predict the crop yield, below are the data sources that we have:

1. **Crop yield data:**
   It consists of recorded crop yield data provided by provincial crop insurance agencies of Saskatchewan, and Manitoba. Originally, this data was obtained at quarter section level and now have been rescaled to township level using geostatistical techniques.

2. **Earth Observation predictor variables:**
   This consists of satellite derived predictor variables which include:
   a. Normalized Difference Vegetation Index (NDVI) from MODIS satellite platform.
   b. Surface soil moisture from active and passive microwave sensors
   c. The evaporative stress index from thermal-optical data
   d. Agro-climate variables derived from station-based weather observation inputs

3. **High resolution modeled weather data**
   This is obtained from the Canadian Meteorological at 2.5km and 10km of radius. These data sets have been aggregated at township level for at least 20 years.

   (Note this dataset is still in processing stage and may time take to work for in terms of data modelling standpoint)

   *****Note: The data is still in processing. Not sure if it will be available for us to use in our study period***

All the above datasets are to be joined to obtain a consolidated table consisting of the predictor variables and the yield as the response variable at a township ID and timestamp level.

Deliverables:

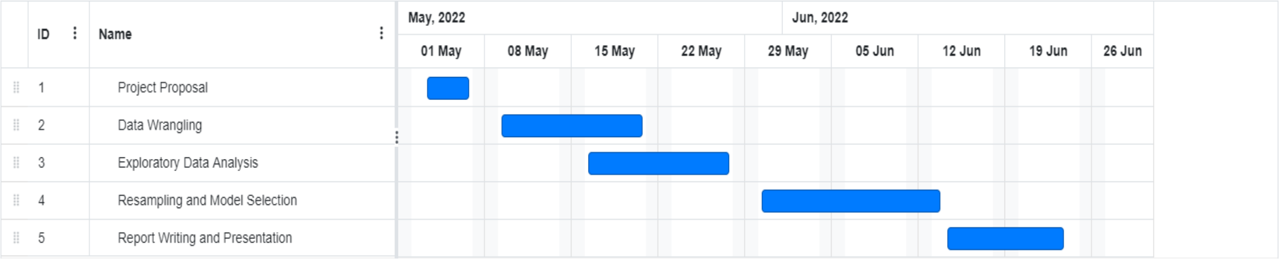
The outcome of the project would involve:

1. Robust crop yield prediction models from ML algorithms using all the climate and earth observation input as they evolve during the growing season will be tested and recommended for the study area.

2. Tabular data containing the predicted crop yield for the test year.

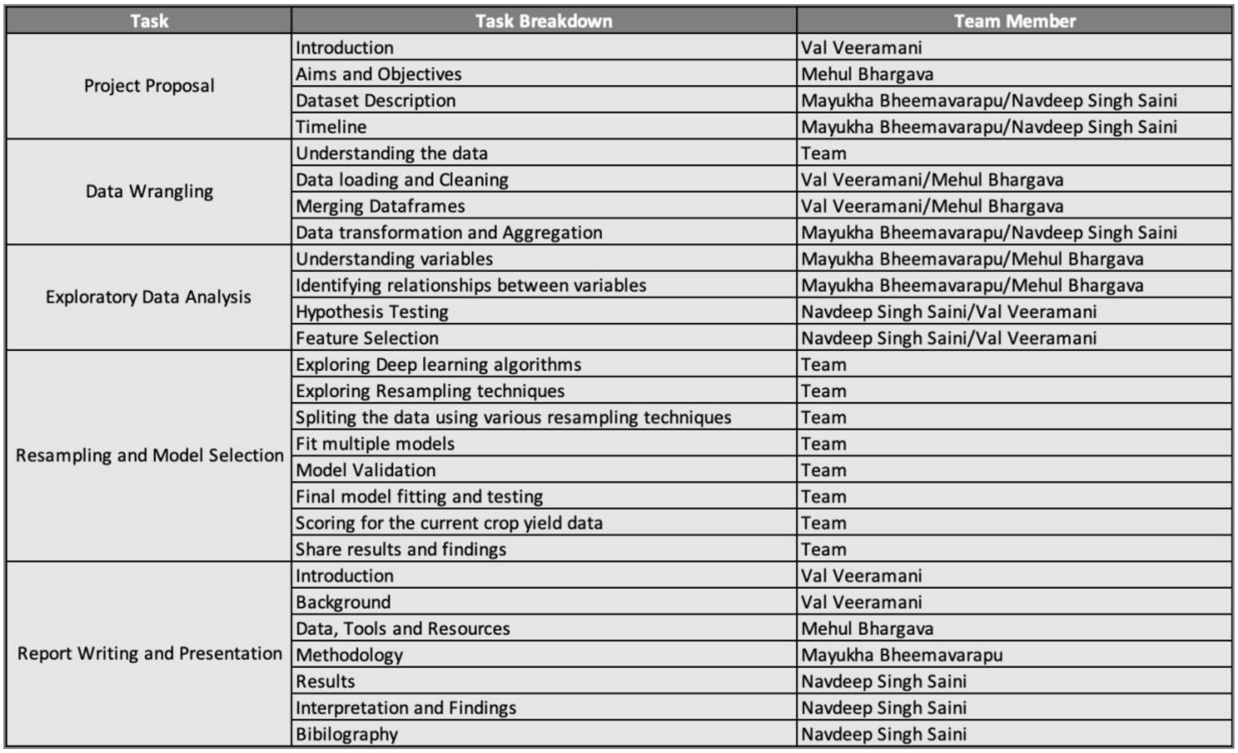
3. Model performance indicators generated using testing data sets.

Timeline:



ID	Name
1	Project Proposal
2	Data Wrangling
3	Exploratory Data Analysis
4	Resampling and Model Selection
5	Report Writing and Presentation

Task Breakdown:



Task	Task Breakdown	Team Member
Project Proposal	Introduction	Val Veeramani
	Aims and Objectives	Mehul Bhargava
	Dataset Description	Mayukha Bheemavarapu/Navdeep Singh Saini
	Timeline	Mayukha Bheemavarapu/Navdeep Singh Saini
Data Wrangling	Understanding the data	Team
	Data loading and Cleaning	Val Veeramani/Mehul Bhargava
	Merging Dataframes	Val Veeramani/Mehul Bhargava
	Data transformation and Aggregation	Mayukha Bheemavarapu/Navdeep Singh Saini
Exploratory Data Analysis	Understanding variables	Mayukha Bheemavarapu/Mehul Bhargava
	Identifying relationships between variables	Mayukha Bheemavarapu/Mehul Bhargava
	Hypothesis Testing	Navdeep Singh Saini/Val Veeramani
	Feature Selection	Navdeep Singh Saini/Val Veeramani
Resampling and Model Selection	Exploring Deep learning algorithms	Team
	Exploring Resampling techniques	Team
	Spliting the data using various resampling techniques	Team
	Fit multiple models	Team
	Model Validation	Team
	Final model fitting and testing	Team
	Scoring for the current crop yield data	Team
	Share results and findings	Team
Report Writing and Presentation	Introduction	Val Veeramani
	Background	Val Veeramani
	Data, Tools and Resources	Mehul Bhargava
	Methodology	Mayukha Bheemavarapu
	Results	Navdeep Singh Saini
	Interpretation and Findings	Navdeep Singh Saini
	Bibilography	Navdeep Singh Saini

References:

Luca Sartore, Arthur N. Rosales, David M. Johnson, Clifford H. Spiegelman, Assessing machine leaning algorithms on crop yield forecasts using functional covariates derived from remotely sensed data, Computers and Electronics in Agriculture, Volume 194, 2022, 106704, ISSN 0168-1699.

Huiren Tian, Pengxin Wang, Kevin Tansey, Jingqi Zhang, Shuyu Zhang, Hongmei Li,
An LSTM neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the Guanzhong Plain, PR China, Agricultural and Forest Meteorology, Volume 310, 2021, 108629, ISSN 0168-1923.

Saeed, Umer & Dempewolf, Jan & Becker-Reshef, Inbal & Khan, Ahmad & Ahmad, Ashfaq & Wajid, Syed. (2017). Forecasting wheat yield from weather data and MODIS NDVI using Random Forests for Punjab province, Pakistan. International Journal of Remote Sensing. 38. 4831-4854. 10.1080/01431161.2017.1323282.

Aston Chipanshi, Yinsuo Zhang, Louis Kouadio, Nathaniel Newlands, Andrew Davidson, Harvey Hill, Richard Warren, Budong Qian, Bahram Daneshfar, Frederic Bedard, Gordon Reichert, Evaluation of the Integrated Canadian Crop Yield Forecaster (ICCYF) model for in-season prediction of crop yield across the Canadian agricultural landscape, Agricultural and Forest Meteorology, Volume 206, 2015, Pages 137-150, ISSN 0168-1923