Corn Growth Prediction for the Upcoming Season in Burkina Faso.

ZINA Lacina ,[1] SUN Yi[2,+]

Department of Informations Systems, Kobe Institute of Computing

Abstract. The economy of many African countries depends heavily on agriculture, forestry and livestock farming, as well as the exploitation of mineral resources. In Burkina Faso, Traditional cereals, such as sorghum and millet, dominate food consumption and expenditure of rural households, while urban households prefer rice and maize. However most farmers today are smallholder or subsistence farmers who grow crops and rear animals just to feed themselves and their families. The lack of information remains the number one problem facing most scale farmers today. Nowadays there are many available platforms that we can retrieve information such as weather information, soil information, yield information that can be used to build a strong prediction platform. It is from there that we came out with our prediction idea based on Corn and extended to other crops in the future work. The solution is a web application and mobile application that will be deployed on the AWS cloud provider. The farmers can install or access from their smartphone or laptop, and based on their geo position the application provides information about the upcoming season. The application is a microservice application combines with machine learning tool such as Scikit Learn.
To put in place this solution we use many procedures (extract, clean and store the dataset), technologies (docker, cloud), programming languages (Flash, Python) and classification methods (Adaboost, Random Forests, OLS) to make for the end users accurate predictions. Historic weather conditions were downloaded using API services from the Darksky.net from 2005 to the current day and then combined with the data of corn for each province since the same interval date coming from multiple sources AGRA, ministry of Agriculture, this combined output is used for the training data. The prediction is for each coming agriculture season in the country.

Keywords: Customer Productivity, Data Science; Data Analysis, Cloud, JSON, Dataset, Data Warehouse, Docker, Docker image, Machine Learning, Scikit Learn Adaboost, Random Forest, Ordinary Least Squares (OLS).

1. Introduction

With more than 60% of its 1.166 billion people living in rural areas, Africa's economy is inherently dependent on agriculture, more than 32% of the continent's gross domestic product comes from the sector [1]. However, agricultural productivity still remains far from developed world standards. Over 90% of agriculture depends on rainfall (Weather) and the influence of market situation. The techniques used to choose the right crops to grow for the year is still unassisted and it is randomly based on the current market situation. Agriculture in Africa experiences basic infrastructural problems such as access to markets and financing facilities. The access to the markets to sell the product is one of the biggest issues in Africa, due to bad transportation infrastructure and non-motorable roads. Farm products don't get to the market on time, preservation facilities are also not available to prevent rot, these lead to a reduction of productivity, thereby leading to an increase in poverty. Agriculture employs 65 percent of Africa's labor force and accounts for 32 percent of gross domestic product [2], but farmers are the poorest due to the problems identified above.

Corn is a global crop which is grown in virtually every country of the world. Countries depend on this very important crop to sustain the local population and foreign exchange earning. Growing corn requires

---

[+]Corresponding author. Tel: +81-78-262-7715
Email address: sun@kic.ac.jp

some basic input such as a minimum level of water, sunshine, favorable weather condition, etc. In recent time, climate change has been one of the most discussed topics on challenges affecting every aspect of our daily life. Changes in climate are adversely affecting the weather of different regions of the world. In Burkina Faso, farmers face the problems of inconsistent rain and dry weather seasons. they are daunted with problems of predicting when these seasons will come around from their experiential knowledge [3].

Governments train different professionals in different domains to help the farmers increase their productivity, but the fundamental problems are still re-occurring because these professionals do not have the necessary tools to help the farmers. Nowadays most of the trained professionals are more concerned with the current government policies other than the benefits of the farmers. For example, the government focus on the product that can only contribute to their exportation balance, and this strategy does not take into account the situation of the demand in the market nor the weather forecast.

Many other projects were developed by **NGOs**[2] to help farmers to have a good harvest funded by the World Bank and others. Most of these solutions focus on productivity, but the population is still facing hunger and farmers are poor. Why is this so? Nowadays, many organizations provide funding for data collection to further research on the factors responsible for these problems [4]. One of the most important datasets that will be useful for the research is the available dataset of AGRA (Growing Africa's Agriculture) based on crops production statistics per year for each province.

The main purpose of this research is to provide a user-friendly web platform to farmers, whereby the farmer only needs to provide the farm location, the platform will provide a prediction and accurate statistics of the cultivation of corn for the coming season. The platform is divided into two parts such as the machine learning side and the Web API side.

2. Related Works

2.1. RiceAdice
The organization Africa rice provided a mobile application called RiceAdvice, an Android-based decision support tool for providing farmers with pre-season field-specific management guidelines for rice production systems in Africa. The guidelines include target yield, nutrient management, crop calendar, and good agricultural practices.

This application is an ordinary application based on the passed data of the farmers. To provide the pre-season guideline, the application will collect the farmer personal condition such as rice growing condition, typical practices, expected sowing date, fertilizer availability, and market price. Based on these data and also on their available budget or desired production level the application provides personalized advice including a crop calendar, fertilizer plan, recommended rice variety and other good agricultural practices [5].

2.2. Plantix
Born of research in the Amazon forest, the Plantix mobile app is helping farmers on three continents quickly identify plant diseases using artificial intelligence. For several years in the Brazilian rainforest, a team of young German researchers studied the emission and mitigation of greenhouse gases due to changing land use. The team's analysis was yielding new knowledge, but the farmers they worked with weren't interested in those findings. They wanted to know how to treat crops being ravaged by pathogens.

PEAT is using two NVIDIA TITAN X GPUs, in scalable link interface mode, in combination with the CUDA parallel computing platform and programming model to not only train the neural network models behind Plantix but also to analyze incoming photos, a process known as inference. Eventually, PEAT hopes to refine its technology so that it can solve plant-disease issues for all types of farmers, not just those growing mass-produced crops [6].

---

[2] *NGO: Non-Governmental Organisation*

3. Proposed Solution

There are a plethora of articles on Deep Learning (DL) or Machine Learning (ML) that cover topics like data gathering, data munging, network/algorithm selection, training, validation, and evaluation. But, one of the challenging problems in today's data science is the deployment of the trained model in production for any consumer-centric organizations or individuals who want to make their solutions reach a wider audience. Most of the time, energy and resources are spent on training the model to achieve the desired results, so allocating additional time and energy to decide on the computational resources to set up the appropriate infrastructure to replicate the model for achieving similar results in a different environment (production) at scale will be a difficult task. Overall, it's a lengthy process that can easily take up months right from the decision to use DL to deploying the model.

This solution gives a comprehensive overview of the entire process of deployment from scratch. The following schema is a global architecture of the solution.
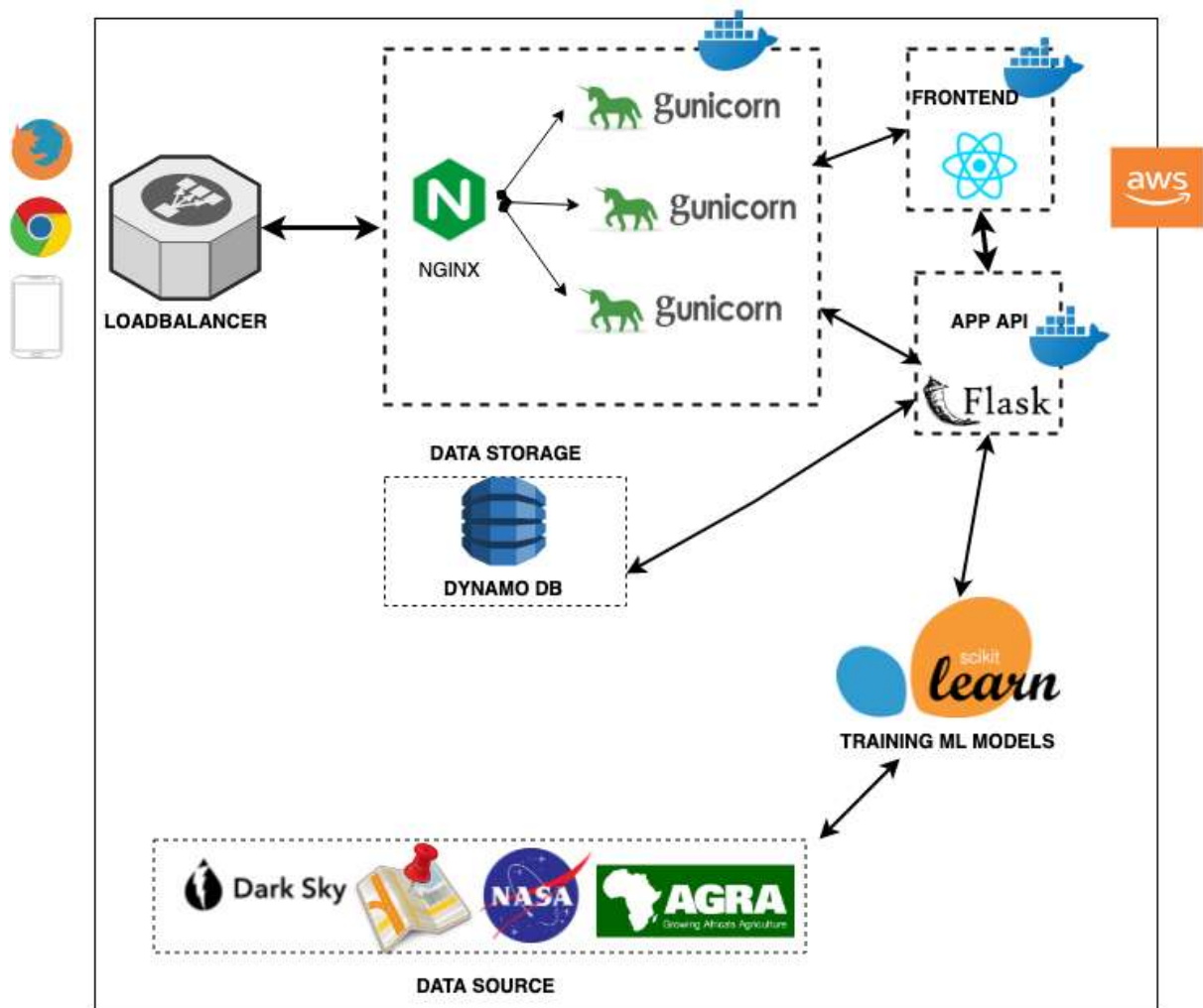

Fig. 1: Schema of the proposed solution.

The project uses many components such as the:
- Frontend part which composes of React application with Nginx as a reverse proxy,
- API service part which collects the input data of the user and sends to the training model, it also stores all the client request in a NoSQL database here the dynamoDB
- The training models use Scikit Learn to train our datasets and use a model such as AdaBoost, random forest or OLS.
- Finally the data source, where we collect all our dataset that help us to make our training model.

3.1. Components

Let's break down the above image that depicts the entire API workflow and understand every component.

- *Docker:* It is a container virtualization technology which behaves similarly to a light-weighted virtual machine. It provides a neat way to isolate an application with its dependencies for later use in any operating system. We can have multiple docker images with different applications running on the same instance but without sharing the same resources. It makes our application more performance by splitting to microservices, and a container can be dedicated to each microservice.
- Nginx is an open-source web server but can also be used as a load balancer. Nginx has a reputation for its high performance and small memory footprint. It can thrive under heavy load by spawning worker processes, each of which can handle thousands of connections. In the image, nginx is local to one server or instance to handle all the requests from the public facing load balancer. An alternative to nginx is Apache HTTP Server. It redirects the traffic of user to different path depends on the request and manage the output of requests whether it API or react page.
- Scikit-learn (formerly scikits.learn) is a free software machine learning library for the Python programming language.[3] It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

3.2. Machine learning/Deep learning

The machine learns section is a computer model to forecast Corn yields in Burkina Faso. The model takes data from a different platform, form the website of Agra Burkina Faso for the yields(corn) data, satellite image to define the type of area, and Darksky.net (weather). Aggregating the result of the satellite images processing with, historic weather, and reported yield of AGRA from 2010 to present date.

3.2.1.   The Different models and libraries used

This section of the project is the intelligent part. It collects, clean, analyze and forecasts the possible results of the upcoming season, then stores in a different location relational database, Redis cluster, and AWS Simple Storage Service (S3). And the API of the application will use this storage to provide the forecasting information to the end users.

- Adaboost

AdaBoost, short for "Adaptive Boosting", is the first practical boosting algorithm proposed by Freund and Schapire in 1996. It focuses on classification problems and aims to convert a set of weak classifiers into a strong one. The final equation for classification can be represented as

$$F(x) = sign(\sum_{m=1}^{M} \theta_m f_m(x)),$$

(1)

where f_m stands for the m_th weak classifier and theta_m is the corresponding weight. It is exactly the weighted combination of M weak classifiers [7].

- Random Forest

Random Forest is a supervised learning algorithm. Like you can already see from its name, it creates a forest and makes it somehow random. The "forest" it builds, is an ensemble of Decision Trees, most of the time trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result [8].

"To say it in simple words: Random forest builds multiple decision trees and merge them together to get a more accurate and stable prediction."
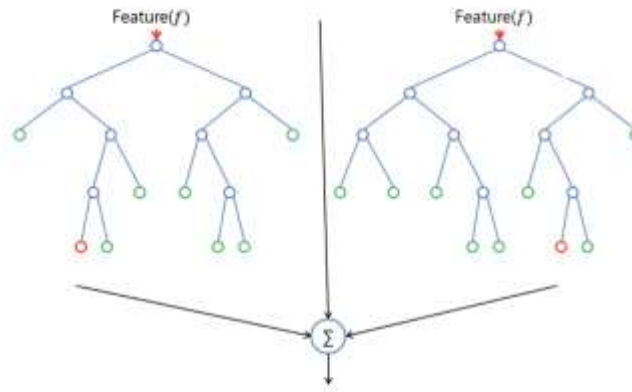
Fig. 2: Random Forest merges trees.

- OLS

Ordinary Least Squares regression (OLS) is more commonly named linear regression (simple or multiple depending on the number of explanatory variables). In the case of a model with p explanatory variables, the OLS regression model writes: $Y = \beta_0 + \Sigma_{j=1..p} \beta_j X_j + \varepsilon$ where Y is the dependent variable, β_0, is the intercept of the model, X j corresponds to the jth explanatory variable of the model (j= 1 to p), and e is the random error with expectation 0 and variance σ^2.

In the case where there are n observations, the estimation of the predicted value of the dependent variable Y for the ith observation is given by: $y_i = \beta_0 + \Sigma_{j=1..p} \beta_j X_{ij}$. The OLS method corresponds to minimizing the sum of square differences between the observed and predicted values. This minimization leads to the following estimators of the parameters of the model: $[\beta = (X'DX)^{-1} X' Dy$ $\sigma^2 = 1/(W - p^*)$ $\Sigma_{i=1..n} w_i(y_i - y_i)]$ where β is the vector of the estimators of the β_i parameters, X is the matrix of the explanatory variables preceded by a vector of 1s, y is the vector of the n observed values of the dependent variable, p^* is the number of explanatory variables to which we add 1 if the intercept is not fixed, w_i is the weight of the ith observation, and W is the sum of the w_i weights, and D is a matrix with the w_i weights on its diagonal. The vector of the predicted values can be written as follows: $y = X (X' DX)^{-1} X'Dy$.

3.2.2. Data collection

Data for the Application comes from a variety of places. Historic weather conditions were downloaded from the Darksky.net (http://www.Darksky.net) servers. Darksky was chosen as it is free (up to 1000 calls per day), and can also handle the task of predicting weather well into the future. Weather data is downloaded in advance (data back to 2010 is in data/weather), the only processing of data was to drop unused columns during the download, in an effort to reduce file size. Additionally, the daily weather data is converted to yearly averages so it can be used with the yield data. YIELDS: The AGRA data of Burkina Faso are collected by the support of an AGRA member, the data retrieved is from the year of 2010. This data is automatically collected by the model for cleaning and combining with the weather data. Since Corn was the focus, only the Corn data was used in the model. Note that all country data gets downloaded so that in the future the model can be easily extended to other countries.

3.2.3. Model and Performance

Three models were run on the data, Ordinary Least Squares, Random Forests, and Adaboost. The primary selected model was Adaboost, as it provided the best **RMSE** (+/- 12.8 bushels/acre on the test set). Data is loaded from the combined file and split into train and test sets. For training mode, the model discards all information not within the 2010 to 2019 date range. The model initially way over-fit on the train data, and after reviewing in-stage Adaboost scoring, the model was 'tuned' to remove overfitting as much as possible. Resulting RMSE fit was 12.8 bushels/acre. Not bad given the wide range of things farms can do to their fields in the course of a year. The following chart shows the model predicted yield (green line) and actual
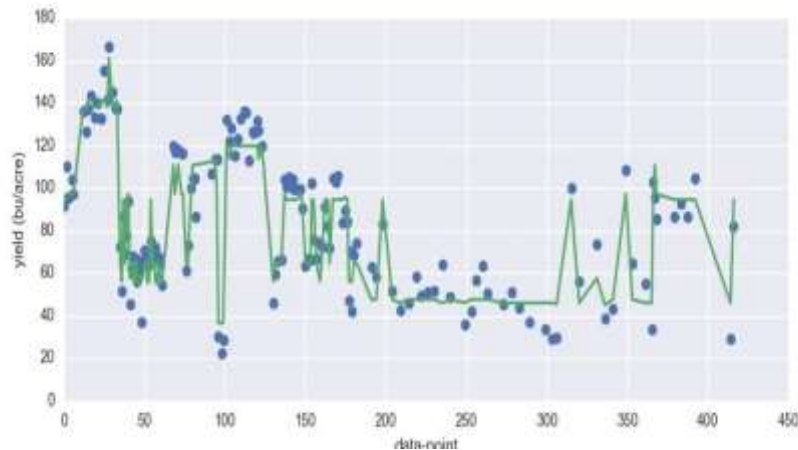
yield (blue dots).

Fig. 3: Yield Predicted and the current yield.

3.3. The output

Since the prediction model based on Scikit-learn is ready, we combined the model with le python framework Flask to build our API. The API is a RestFul API that takes as parameters the current position of the user, after receiving the parameters it will trigger the prediction model.

For the execution of our API, we have chosen Gunicorn to interpret our code. Gunicorn is a good choice if you have built the APIs using Flask. An example command to run the Gunicorn web server. To test the effectivity of the application we build client application which consumes our API Restful. Based on react JS. React is one of the famous and nice javascript framework nowadays. This reactt application is the final output of the research.

4. Conclusion

The predictions application are based on microservices technologies. With two big part, the machine learning part we use three different prediction models and try to compare each other, then serve the best prediction. The collected information from our data providers (Darksky, AGRA, Google API, NASA images) are stored in different resources on the cloud, they are used to train our model. And the flask API takes as input the location of the farmer and return the prediction of the percentage of success of the corn on that location. The application can be improved by adding an image processing on the machine learning side, that will provide a piece of accurate information about the land position.

5. Acknowledgement

6. References

[1] Biosciences for farming in Africa, "Agriculture in Africa: potential versus reality", http://b4fa.org/agriculture-africa-potential-versus-realit.

[2] The World Bank, "Fact Sheet: The World Bank and Agriculture in Africa", available on http://web.worldbank.org/WBSITE/EXTERNAL/COUNTRIES/ AFRICAEXT/0,,contentMDK:21935583~pagePK :146736~piPK:146830~theSitePK:258644,00.html.

[3] John F. Morton, 2007. The impact of climate change on smallholder and subsistence agriculture.PNAS December 11, 2007 104 (50) 19680-19685; https://doi.org/10.1073/pnas.0701855104.

[4] Africa - FAO http://www.fao.org/3/W8376E/w8376e03.htm by D. E. Tempelman

[5] Africa Rice. Tools for improving rice value chains in Africa. https://www.riceadvice.info/en/

[6] PEAT Technologies. PLANTIX. Digital Green Thumb: How AI Helps Farmers Combat Plant Disease. https://blogs.nvidia.com/blog/2016/12/13/ai-fights-plant-disease/

[7] Michael Sprenger Institute for Atmospheric and Climate Science, ETH Zürich, Zurich, Switzerland. Nowcasting Foehn Wind Events Using the AdaBoost Machine Learning Algorithm.https://doi.org/10.1175/WAF-D-16-0208.1, May 2017.

[8] Gregory R. Herman and Russ S. Schumacher Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado. Money Doesn't Grow on Trees, but Forecasts Do: Forecasting Extreme Precipitation with Random Forests. https://doi.org/10.1175/MWR-D-17-0250.1, May 2018.