Capstone Project - Group 15
Predicting Crop Production in Nigeria

Chidi Oshakwuni-Austin
Team Lead

Ezealaji Lucky Damain
Data Preparation

Folake Fadeyi
Data Visualisation

Dumale Igbara
Modelling

Okoye Nzubechukwu Moshe
Modelling

Obioma Tony Chikere
Reporting

Olajide Oluwapelumi
Reporting

Abstract **- Previous research has shown that food shortage is one of the top threats across many continents. The objective of this project is to examine the relationship between food production and population growth in Nigeria to determine if its production rate can match its fast rising population. This is a case of supervised learning, the data on population growth and food production is continuous. Among three regression models, we chose the Support Vector Regression as our final predictive model with a 97% accuracy and least error. The paper explains the exploratory data analysis, methodology and implementation of the algorithm.**

1. INTRODUCTION

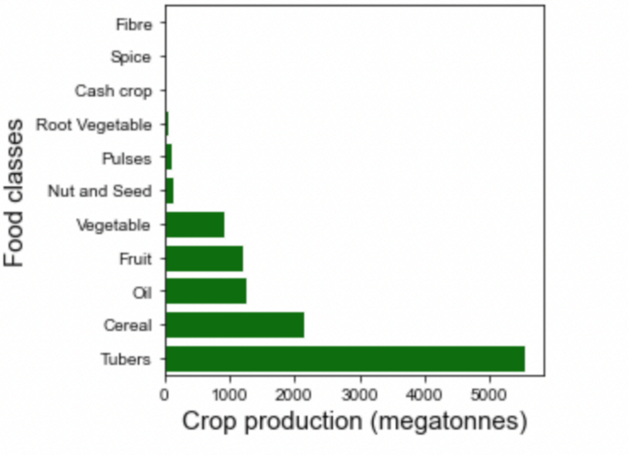
Nigeria is most populous black nation in the world. It has a population count ranging between 200 - 220 Million people with over 250 ethnic groups, occupying 923,768km² of land located in the West part of Africa. Nigeria is known for its rich economy and a great exporter of oil, entertainment, fashion, tech and other creative sectors. However, its agricultural sector has awfully suffered despite having vast natural resources and fertile lands. This report analyzes the population growth from 1961, the year after Nigeria's independence, to 2019. The prediction of crop production rate in Nigeria is complex and requires an in-depth analysis of various factors/variables such as policies, importation, insecurity, flooding, climate change etc. As a result of the limited data in Nigeria, this report focuses on predicting crop production with respect to population growth, using the best indicator.

2. DATASET, FEATURES AND PREPARATION

The crop production dataset was web scraped from 'Kaggle.com' uploaded by Food and Agriculture Organization of the United Nations (FAO). The dataset consists of different types of crops ranging from cereals, citrus, fibre crops, fruits, oil crops, pulses, roots and tubers etc and their yearly harvested area, yield and production across different countries. For the purpose of this project, we were only interested in the production quantity of the agricultural products produced in Nigeria. So, a subset for only Nigeria was extracted from the main data and to do this, some data science tools were utilized. Data cleaning was carried out on the extracted data after discovering some irregularities such as missing values and unwanted columns. The unwanted columns were dropped and the rows containing the missing values were dropped using numpy and pandas packages.

Further analysis required the use of the population data of Nigeria. The data was also web scraped from 'kaggle.com'. The population dataset contained the population statistics of multiple countries of the world from 1950 to 2100. A subset for Nigeria's population between 1981 and 2019 was extracted from the original dataset, and also the population density. The extracted data was cleaned, like the production data, to remove rows containing the missing values and also get rid of unwanted columns using numpy and pandas data science tools. In order to examine the rate of crop production against the population growth, a final dataset was made by combining total food production quantities to show the summarisation of production for each year, which was used to plot against the total population growth in Nigeria.

An obvious detail in the production data showed that some classes of crops were in higher production than others, which may be a factor of the soil type or climate condition of the country. To have a clearer view of the various crops and their quantities, we created a second dataset to target a subset of yearly production of the each crop in Nigeria, showing the different categories of the food and their metric quantities with a bar plot.

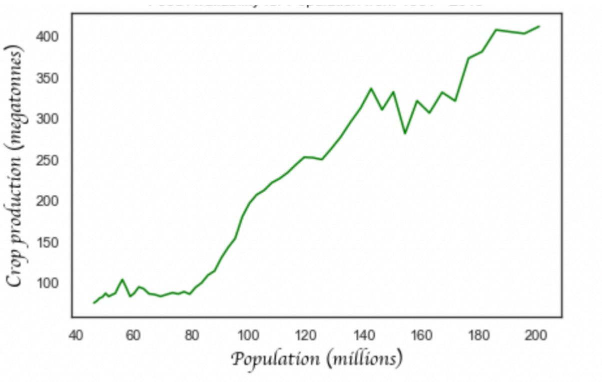


PLOT I: *Plot showing the total count of different classes of food produced from 1961 - 2019*

The bar plot indicates that there has been a high production of 'tuber crop' production in Nigeria. Total production in tonnes of tubers totals 5548mgt compared to cereal at 2156mgt, which is half the production count of tubers. Other classes like fibre, spice and cash-crops seem to be grown in very little quantity at 0.04mgt, 0.24mgt and 6.60mgt respectively.

3. MODELLING

After a round up of the initial explanatory data analysis, we moved to the supervised learning phase.



PLOT II : *Plot showing the relationship between total yearly crop production and population*

The graph in the above plot shows a fairly strong linear relationship between the two continuous variables, which prompted us consider regression models. We applied different machine learning models to the data: linear regression (LR) , bayesian linear regression (BLR) and support vector regression (SVR). We fed in 'population count' and 'population density' features as our 'x' to predict crop production, out 'y'. The data was split 80-20 for training and testing sets respectively and set at a 'random state' of 42 for all models to ensure same results or numbers when we run the tests.

Linear Regression is a popular algorithm used in forecasting or predicting continuous data before improved versions were introduced. Its cost function finds the best values for intercept and slope to provide the best line for prediction. The linear regression was imported from SciKit Learn package.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$$

Application of the LR model was straightforward, it required no hyper tuning because of its lack of hyper parameters

Our second model, the bayesian linear regression is different type of linear regression that uses probability distributions instead of point estimates to predict. Within conventional statistical techniques, the null hypothesis is always set up to assume no relationship between the variables of interest. This null hypothesis makes sense when one lacks understanding of the relationship between the variables. Bayesian statistics allows us to overcome this constraint, learn from our data and incorporate new knowledge into future predictions. Both model parameters and estimated parameters comes from probability distributions. Thus the optimized model:

$$P(\beta|y, X) = \frac{P(y|\beta, X) * P(\beta|X)}{P(y|X)}$$

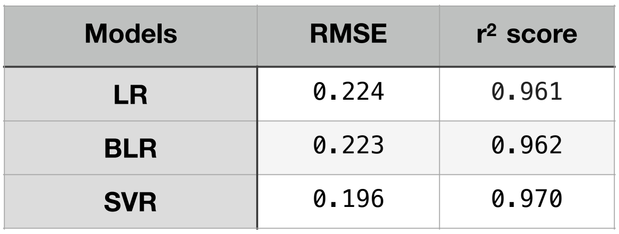
Similar to the LR model, the bayesian regression model has no hyper parameters to tune. Another very straightforward algorithm, making us consider a third model with hyper parameters.

After a detailed research and consultation of different articles on SVR modelling, the SVR model was applied on two features of the population dataset and the crop production dataset to model the relationship between them and also predict the rate at which production matches the growing population. The SVR uses classification algorithm to predict regression values and by so doing, acknowledges the non-linearities in the dataset unlike the our other models. We applied SVR to our problem using the algorithm implemented in Sci-Kit Learn and proceeded to test the model on multiple kernels including linear, Gaussian (radial basis function) and polynomial.

We noticed an interesting difference in the computation time with the polynomial kernel and significant overfitting with the Gaussian kernel. Removing features from the data matrix removed the overfitting problem with the Gaussian kernel, but brought performance down to approximately match the linear kernel. In order to get results from the polynomial kernel within a reasonable time frame, we considered reducing the data matrix, additionally, after taking these steps the polynomial-kernel SVR model was much more sensitive to the random splitting of training and test data than the other kernels.

The modelling was concluded by calculating the r² scores and RMSE scores for accuracy; and errors respectively, then visualized the test values and predicted values of these models.

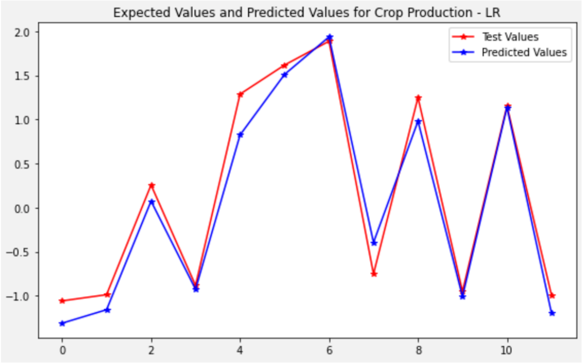
4. RESULTS



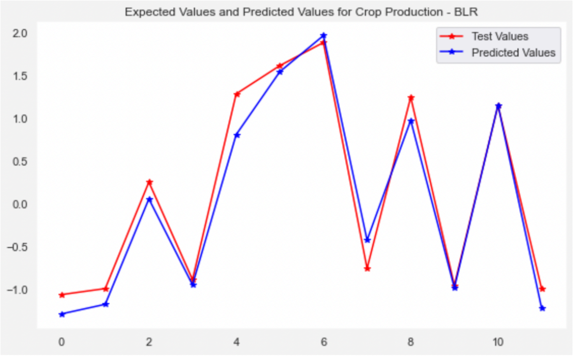
Models	RMSE	r^2 score
LR	0.224	0.961
BLR	0.223	0.962
SVR	0.196	0.970

TABLE 1: *RMSE and r² scores of models.*

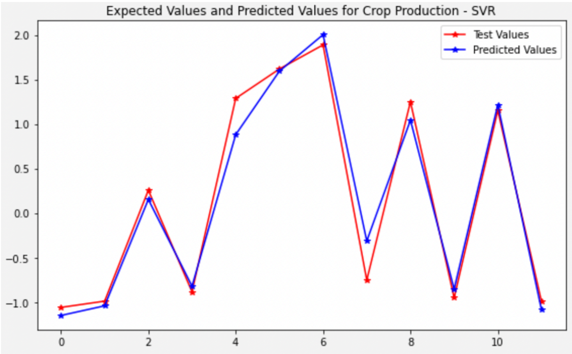
Comparing results, the table above shows the r² scores and RMSE of all three models. These models were built on yearly predictions, meaning that each prediction is the expected total crop production one year into the future. The SVR model provided the best fit with an r² score of 97% because of its ability to take the nonlinearities into its function making it superior to the linear regression and bayesian regression with accuracy of 96.1% and 96.2% respectively. It is interesting to note that the RMSE scores are also in sync with the r2 scores. The SVR model register the least errors of the three models.



PLOT III: *Crop production test values and predicted values for Linear Regression model*



PLOT IV: *Crop production test values and predicted values for Bayesian Regression model*

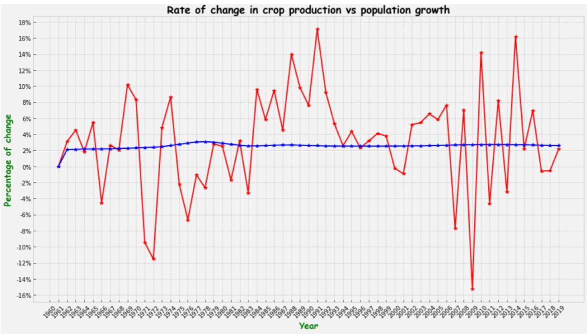


PLOT V: *Crop production test values and predicted values for Support Vector Regression model*

The plots above display an overlay of crop production actual values with the predicted values for each model. Comparing all three models, we see that the SVR lines have the best fit meaning they are closer to the actual values. These three metrics show that the SVR model has the best predictive power, greater than the other two.

5. ANALYSIS

To further understand our data, we examined how the selected features in our final datasets were related. Using a correlation heat map, we ascertained that population count and crop production are highly correlated, which means that Nigeria's crop production is highly dependent on its population. The yearly change in population grew at a steady increase of 2% while the yearly change in crop production varied all through the years, this shows that that there are other important features that affect crop production as stated earlier in this report.



PLOT III: *Plot shewing change in growth of population and production across the years*

Choosing our preferred model amongst the three was easy and glaring. Under the condition of root mean square error (RMSE) and the r^2 score, we chose SVR to be our predictive model.

6. CONCLUSION

From the research, we were able to produce highly accurate results using all three models with SVR showing best results. The SVR model is known for being the best model for forecasting or predicting. Our model can be applied in the Agricultural sector as a 'benchmark' to set a food production target for Nigeria with respect to its fast rising population in order to detect and remedy food scarcity, a leading threat in the world especially in Africa, ahead of time. We believe for this purpose, there is room for improvement on the model to factor in other features that also affect crop production for more efficient results. The unavailability of data for these features in Nigeria contributed to our limitations in this work.

REFERENCES

- Olimar E. Maisonet-Guzman. (18 July 2011). *Food Security and Population Growth in the 21st Century.* E-International Relations. https://www.e-ir.info/2011/07/18/food-security-and-population-growth-in-the-21st-century/

- Dennis Erezi. (29 May 2020). *Nigeria Lacks Protein*. The Guardian News https://guardian.ng/news/report-says-nigerians-lack-protein-as-rice-eba-amala-top-food-consumption-list/

- Saul Dobilas. *Support Vector Regression (SVR)—One of the Most Flexible Yet Robust Prediction Algorithms.* Towards Data Science https://towardsdatascience.com/support-vector-regression-svr-one-of-the-most-flexible-yet-robust-prediction-algorithms-4d25fbdaca60

- Will Koehrsen. *Introduction to Bayesian Linear Regression.* Towards Data Science https://medium.com/towards-data-science/introduction-to-bayesian-linear-regression-e66e60791ea7

- https://stats.stackexchange.com/questions/220158/svm-training-and-testing-error-interpretation