Prediction of Economic Factors Using Machine Learning Techniques

MLPC V : MACHINE LEARNING PROJET COMPETITION 2023_2024

Departement of Mathematics Physical Sciences (MPS)

Author 1: **Balbino Tchoutzine**; tchouzine@gmail.com

Author 2: **Jacquy Ngonga**; junsts719@gmail.com

1. Abstract

In a world where the economy dictates the pace of progress and the well-being of nations, the ability to predict economic indicators becomes a powerful tool for strategic decision-making. The project presented aims to address the problem of economic unpredictability in Cameroon, focusing on key indicators such as Gross Domestic Product (GDP), GDP per capita, and inflation rate. By leveraging machine learning techniques, this work proposes a predictive model tailored to the Cameroonian context, capable of providing accurate and reliable forecasts that can serve as a foundation for informed economic decisions.

2. The Motivation Behind the Problem

The motivation behind this project stems from a tangible reality: the Cameroonian population is often caught off guard by sudden fluctuations in the prices of basic commodities, particularly oil. This volatility has a direct impact on daily life and can lead to social instability. By anticipating these variations through the prediction of economic indicators, it is possible to prepare society for these changes and to implement proactive policies to mitigate negative effects. This project aims not only to predict the economic future but also to provide a tool for adaptation and preparation, thus strengthening Cameroon's economic resilience in the face of global challenges.

The Importance of Machine Learning in Economic Forecasting

Machine learning has become an indispensable tool in the field of economic forecasting. Its ability to analyze vast datasets and identify complex patterns allows for more accurate predictions of economic trends and outcomes. By leveraging advanced algorithms, machine learning can uncover insights that traditional statistical methods may overlook, providing a deeper understanding of economic dynamics. This technological advancement is particularly valuable in today's rapidly changing economic landscape, where traditional forecasting models may struggle to keep pace. As such, machine learning represents a significant step forward in the pursuit of precise and reliable economic forecasts.

Why Use Machine Learning ?

- ✓ Complex Modeling: Economic relationships are seldom linear. Machine Learning (ML) allows for the modeling of complex interactions among various variables, which is crucial for understanding economic trends.
- ✓ Abundant Data: We now have access to a massive amount of economic data, ranging from interest rates to exports. ML can extract useful information from these large datasets.
- ✓ Adaptability: The economy is dynamic. ML can quickly adapt to changes, providing real-time forecasts.

Caption: A caricature depicting the reaction of gas stations and citizens following the price increase.

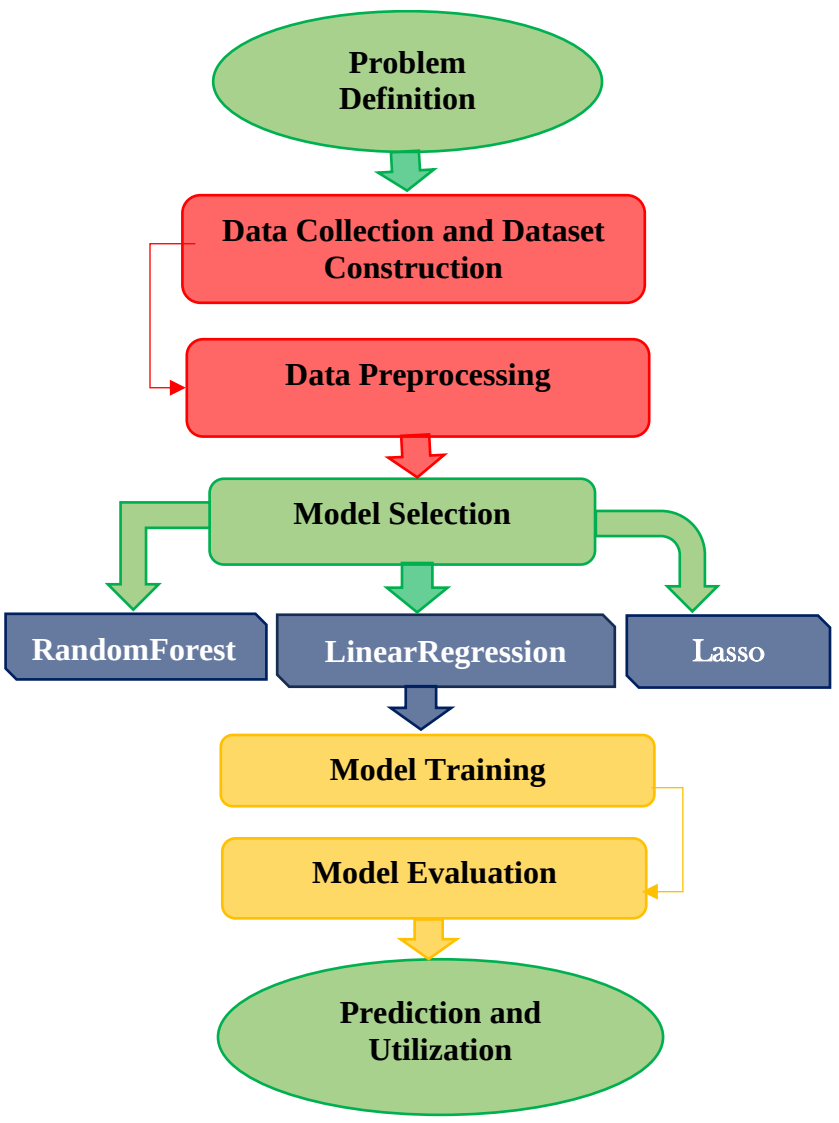
GDP per Capita and Preparedness

GDP per Capita as an Indicator of Average Wealth: A Key Metric with Limitations in Reflecting Population Preparedness for Price Inflation:

- **Purchasing Power:** If prices rise more rapidly than GDP per capita, purchasing power diminishes, impacting the quality of life.
- **Financial Education:** Preparedness for inflation also hinges on financial education. Public policies must bolster this education to assist the populace in navigating economic challenges.

In conclusion, Machine Learning is a potent tool for anticipating economic fluctuations; however, it must be supplemented by informed policies to prepare the populace for impending economic challenges.

3.Méthodologie de Résolution du Problème



A. Problem Definition

- **Specific Problem**: Specific Problem The particular issue we aim to address is the economic unpredictability in Cameroon. Our objective is to forecast the fluctuations of economic indicators such as the Gross Domestic Product (GDP), GDP per capita, and the inflation rate.
- **Goals to Achieve**: Objectives to Achieve Our aim is twofold. Firstly, we strive to provide accurate and reliable forecasts for these economic indicators. Secondly, we intend to use these forecasts as a foundation for informed decision-making, thereby contributing to the enhancement of Cameroon's economic resilience in the face of global challenges

B. Data Collection and Dataset Construction

Within the Scope of Our Gross Domestic Product (GDP) Forecasting Project In the context of our project aimed at predicting the Gross Domestic Product (GDP), we were required to compile a robust dataset. This process, while essential, proved to be complex and demanding. Here are the parameters we have considered:

- **Industrial Production Outcomes (oil, coffee, cocoa, gas, etc.):** We have examined the performance of industrial sectors in Cameroon. These data have enabled us to understand production trends and assess their impact on the GDP.
- **Population Growth:** Demographic growth is a key factor. We have gathered data on the population's evolution over the years, as it directly influen ces the demand for goods and services.
- **Inflation Rate:** Inflation affects the real value of the GDP. We have closely monitored the variations in the inflation rate to adjust our forecasts.
- **Historical GDP Data:** To train our models, we have utilized past GDP values. These data have allowed us to establish relationships and predict future GDP levels.

Challenges Encountered: Data collection was not without its difficulties. Despite nights of research and exploration on national and global web pages, such as the National Hydrocarbons Corporation **(SNH),** the Bank of Central African States **(BEAC)**, the National Institute of Statistics **(INS)**, and the **World Bank**, certain parameter values proved to be unavailable. Faced with this constraint, we had to make rigorous decisions. We discarded corrupted data during the cleaning process to ensure the quality of our dataset and prevent model corruption.

In summary, data collection is a crucial step in any machine learning project. The quality of the data directly impacts the accuracy of our models and, consequently, the reliability of our economic forecasts.

C. Data Preprocessing

Data Collection and Preprocessing Data collection is an essential step, yet another critical stage is making the data usable. Data gathered from various sources may be disorganized and contain numerous null values, as experienced in our project. They may also hold invalid data and non-relevant information. The fundamental steps of data preprocessing include cleaning these data, replacing them with suitable or approximate data, eliminating null and missing values, and substituting them with fixed alternative values. During this phase, we removed missing entries, standardized the typing of different values, and organized the data coherently, ensuring that the model could interpret them accurately.

Correlation Analysis Among Various Parameters

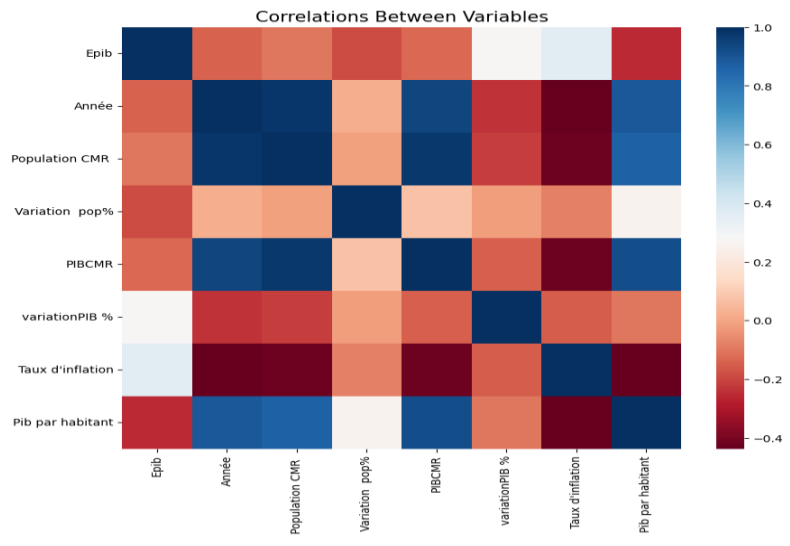
The Importance of the **Correlation Graph** in Our GDP Forecasting Project The correlation graph is an indispensable tool in our project for predicting the **Gross Domestic Product (GDP**) in Cameroon. Here's why:

Identification of Relationships: The correlation graph allows us to visualize the relationships between different variables. We can detect positive correlations (when two variables increase together) or negative correlations (when one variable increase while the other decreases).

Variable Selection: By analysing the graph, we can choose the most relevant variables for our model. Strongly correlated variables may be redundant, while weakly correlated variables can be excluded.

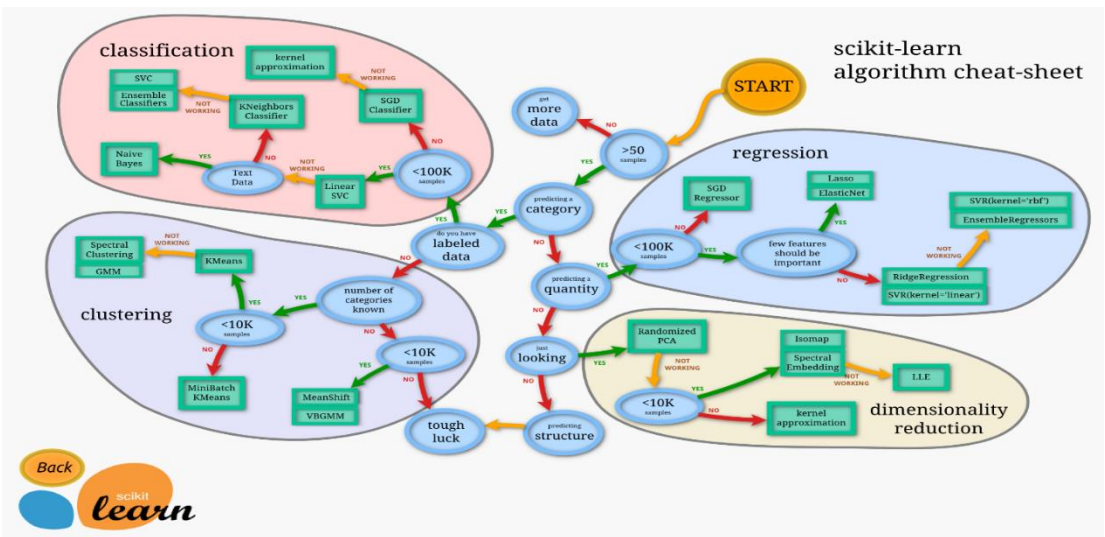
More Accurate Prediction: By using correlated variables, our prediction model can be more precise. For example, if GDP is strongly correlated with industrial production, including this variable will improve our forecasts.

In Summary The correlation graph guides us in the selection of variables and contributes to the robustness of our GDP prediction model.



Correlation Graph between the Various Parameters of Our Data

D. Model Selection



Model Selection Map This map, proposed by Guillaume, a Data Scientist in London, has facilitated the selection of an appropriate model for the project. It allows for the proper selection of a model based on the quantity of data available. However, the choice of the model should not be limited to this aspect alone. It must also be made in accordance with the objective to be achieved.

Why a Regression Model Rather Than a Classification Model?

Choosing Between Regression and Classification Models When selecting between a regression model and a classification model, we must consider the context and specific objectives of our project. Here's why a regression model may be more appropriate in certain scenarios:

Nature of Output Variables :

- **Regression:** Used when the output variable is continuous (for example, GDP, temperature, the price of a house).
- **Classification:** Used when the output variable is discrete and categorical (for example, predicting classes, such as "spam" or "non-spam").

Prediction Objective:

- **Regression:** Aims to predict a precise numerical value. For instance, forecasting a country's GDP.
- **Classification:** Aims to assign a class or label to an observation. For example, categorizing patients as "ill" or "not ill".

Interpretation of Coefficients:

- **Regression:** The coefficients (slopes) have a direct meaning. They indicate the impact of each independent variable on the output variable.
- **Classification:** The coefficients are less interpretable because they are related to class probabilities.

Flexibility :

- **Linear Regression:** It is simple and easy to understand, making it well-suited for scenarios where the relationship between variables is linear.
- **Classification:** It offers more flexibility with algorithms such as random forests, SVMs, etc.

In Conclusion: The choice between regression and classification depends on the context, the nature of the data, and the prediction objectives. In our case of GDP forecasting, regression is appropriate because we are aiming to estimate a continuous value.

Choice of Regression Model

The **linear regression** was selected for our Gross Domestic Product **(GDP)** prediction project for the following reasons:

- ✓ **Linear Relationship:** Linear regression is suitable when a linear relationship between variables is suspected. In our case, we aim to predict GDP based on other factors, and linear regression allows us to model this relationship.
- ✓ **Interpretability:** Linear regression provides coefficients for each variable, enabling us to understand the impact of each factor on GDP. For instance, we can assess the effect of an increase in public spending on GDP.
- ✓ **Simplicity:** Linear regression is a straightforward and easy-to-implement model. It offers us an initial approximation of the relationships between variables without requiring excessive complexity.

In Summary Linear regression is a prudent choice for our GDP forecasting project as it allows us to analyze and model the relationships between variables in a transparent and interpretable manner.

Regression Model Description

Linear regression is a statistical model used to represent the relationship between a dependent variable (the target) and one or more independent variables (the features). Let's delve into the mathematical description of simple linear regression:

a. **The simple linear regression model:**
   - **Model Equation :** [y = a x + b]
     - (Y): Represents the dependent variable (target).
     - (X): Represents the independent variable (feature).
     - (a): Coefficient of the independent variable (slope of the regression line).
     - (b): Intercept term (the value of (Y) when (X) is zero).

b. **Objectif** :
   - Find the coefficients (b) and (a) that minimize the sum of the squares of the residuals (differences between observed and predicted values

In summary, linear regression models a linear relationship between a target variable and a feature, using a regression line to represent this relationship

E. Model Training

Training and Testing

After data processing and model selection, the next step is obviously training and testing. This is where the performance of the algorithm, the quality of the data, and the required output become apparent.

Here's how it works:

Training: We use, by applying Pareto's method, 80% of the data from the vast dataset to train our model. According to it, **20%** of the actions produce **80%** of the results. The training allows the machine to learn and develop the ability to make future predictions.

- **Another:** Learn and develop the ability to make future predictions.
- **Test:** We reserve the remaining 20% of the data for testing. This test set is already labeled with known outputs. We test the model to check if it functions correctly and if it provides the right predictions.
- **Evaluation**: Here we will use the MSE to evaluate errors. The MSE, or **Mean Squared Error** (mean squared error), is a commonly used measure to evaluate the performance of a regression model. It calculates the average of the squares of the differences between the values predicted by the model and the actual values.

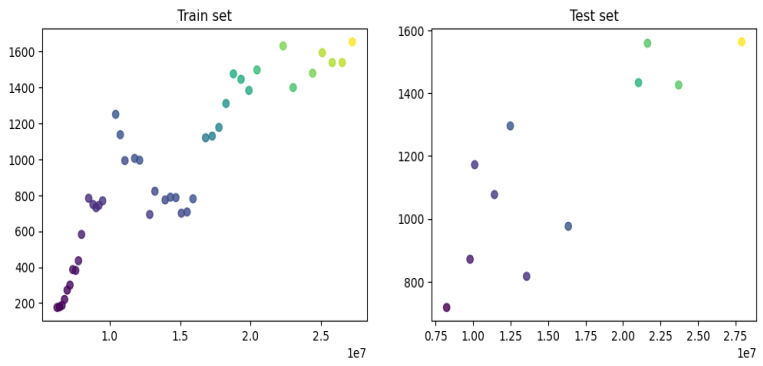
$$j(a,b) = \frac{1}{2m}\sum_{i=1}^{m} (\,\mathbf{f}\,(\mathbf{x(i)}) \;-\; \mathbf{y(i)}\,)^2$$

F. Resultat

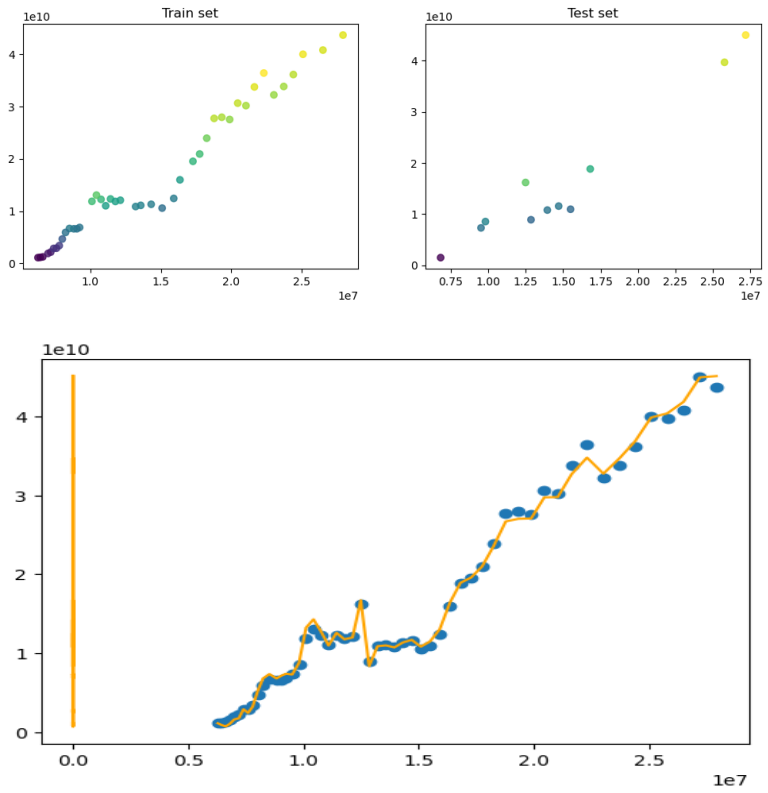
The data was primarily trained and tested using the linear regression model, as an in-depth analysis conducted on the dataset led us to this model. Thanks to the analytical tool (Sklearn Map) proposed by Guillaume, it was not necessary to verify all other regression models. However, for the sake of thoroughness, we examined other models, including Lasso, Ridge, BayesianRidge, and the Random Forest regressor. The goal was also to eliminate any doubt regarding the overfitting of our model. We have limited this justification to the small volume of data, which must always reach a specific threshold, especially for Cameroon, given its recent independence in 1961. It is important to note that it was only a few years after this date that GDP became a relevant economic target and was taken into account in our analyses for Cameroon.

Modeles	**Regression Lineaire**	**Bayesian Ridge**	**Random Forest Regressor**	**Rigide**	**Lasso**
Score 1	0.995471736345961515	0.946098676765206 5	0.995367305492799 6	0.984862a007216231	0.995471736335006 7
GDP and GDP Per Capita					
Score 2	0.984876960239349 3	/	/	/	0.987083587540681 2

Quelques graphiques d'entraînement du modèle.



Comparison of Training and Test Sets Using Scatter Plots (For Both Predictions Made)



Model Evaluation Curve the first prediction

Currently, we have only developed a function in the Jupyter notebook that allows the user to enter parameters or estimates related to the prediction date. In return, this function provides the predictive value of the GDP. This feature has been implemented for both GDP measures.

5. Conclusion and future work

In conclusion, the approach adopted for the prediction of economic indicators in Cameroon illustrates a major advancement in forecast analysis and strategic planning. The use of machine learning techniques, and particularly the linear regression model, has enabled a notable accuracy in predicting economic trends. This technological progress offers an unprecedented opportunity for Cameroon to equip itself with reliable predictive tools, essential for navigating confidently in an uncertain global economic environment.

In the context of our future work, we will develop an application to host this model instead of the simple function, although it returns the expected result, it is not in the layman's environment, thus facilitating access to economic forecasts and enabling more informed decision-making. But before we get there, we will adjust new data related to these economic factors to replace those that have been discarded, with the aim of refining the model and providing better results. Finally, we will look at other economic factors such as the inflation rate, the purchasing power of the Cameroonian consumer or housewife, and even the national debt, depending on the data we will collect.

In the end, this economic prediction project does not merely outline the contours of a potential economic future for Cameroon, but actively commits to forging a resilient economy, capable of meeting current and future challenges, while paving the way for sustainable prosperity for the country.

6. Related Work and Inspirations

In the course of our work, we have explored several other projects of a similar nature, specifically **predicting apartment prices based on various parameters**. No direct relationship exists between these two projects: one aims to predict an economic factor such as GDP, while the other focuses on predicting apartment prices. However, the fact that both projects utilize discrete variables and linear regression models has captured our full attention.

Regarding predicting GDP in Cameroon, no open-source work has been found after several days of research. However, information about the parameters used does exist, creating a distinction between our model and others. Indeed, the parameters typically employed in this field have a more significant impact than ours, especially when we have extensive data availability. I am thinking here of the case of the National Institute of Statistics **(INS).**

Having a wide range of data can lead to the neglect of subtle elements, but they have their role to play. For example, the evolution of the population at first glance does not have a big impact on GDP, but our results and certain curves have proven otherwise. Our project may not be the best in its field, but with **hyperdata directly related to GDP**, taking into account these subtleties, we could have a version that is **200% accurate** before our eyes.

In the context of this project, we have gone through several courses in search of inspiration and to deepen our understanding of the models. Among others, I could mention the course and book by **Guillaume Saint-Cirgue** on 'Learning Machine Learning in a Week.' Do not be misled by the indicated duration, as this course offers a wide range of knowledge, from everything you need to know about Python applied to Machine Learning to the selection and training of prediction models. It is available on his Machine Learnia platform. We also explored the **CS229 courses** '**Machine Learning Introduction 2022**' and '**Weighted Least Squares**, **Logistic Regression**, Newton's Method I 2022.' These courses have allowed us to acquire a general knowledge of Machine Learning, as well as a deep understanding of the linear regression model and error evaluation, which is often a difficult part to find elsewhere. Finally, we participated in various workshops organized by the **AI unit** and the ML training workshops of the **GI club**, led by Professor **Christopher Thron.**

In short, although these courses were not followed to completion, we have gathered a maximum of useful knowledge for our project. Our passion for this field has allowed us to carry the project through, at least for this phase

7. References

[1] https://ins-cameroun.cm/statistiques/
[2] https://donnees.banquemondiale.org/

[3]CS229: Machine Learning (stanford.edu)
[4] https://youtu.be/k_pDh_68K6c

[5] https://youtu.be/Bl4Feh_Mjvo

[6] https://www.beac.int/statistiques/

[7] https://youtu.be/EUD07IiviJg

[8] https://www.afdb.org/fr/countries/central-africa/cameroon/cameroon-economic-outlook

[9] https://nso-cameroon.opendataforafrica.org/apps/data-catalog

[10] Institut National de la Statistique du Cameroun – Des statistiques de qualité pour un meilleur accompagnement du Cameroun vers son émergence (ins-cameroun.cm)

[11] Former MLPC Project : Gender Bias Detection in Job Descript ions; Student Speciality Counselling; Optical Character Recognition for Cameroonians Languages: case Yemba