Billy Kipchirchir Koech

Team Awesome

PROJECT TITLE:

ANALYSIS OF YOUTH EMPLOYMENT IN RURAL AND URBAN AREAS IN TERMS OF SEX AND ECONOMIC ACTIVITY

INTRODUCTION

My name is **Billy Kipchirchir Koech**.

I am a master’s student at the **University of Nairobi.** Pursuing a master’s in **Transportation Engineering.**

I am also pursing an **IBM Data Science Professional Certificate on Coursera.**

WHY I PARTICIPATED IN THE HACKATHON

1. To solve a real world problem that’s closer to home ie youth employment in Africa.
2. Gain experience and expertice handling and analysing raw data.
3. To get the internship position at World Data Lab.

STEPS TAKEN IN DEVELOPING THE PROJECT

STEP 1: LOADING DATA

In this step I loaded all the three files ie Kenya National, Ghana National and Rwanda National to the jupyter notebook. I merged all of them into one.

STEP 2: DATA WRANGLING

In this step I grouped the merged data by indicators and choose the ‘Employment by sex, rural/urban areas and economic acticity (EMP_2EMP_SEX_GEO_ECO_NB)’ indicator for analysis. I cleaned the data and prepared it for analysis. (The data cleaning process is explained in the next slide.)

STEP 3: DATA VISUALIZATION

In this step I develped a dashboard to visualize the data. The dashboard groups the data by country, region and year and displays a bar chart showing the observed values against economic activity grouped by gender.

STEP 4: FEATURE ENGINEERING, MODEL SELECTION AND MODEL REFINEMENT

In this step I one encoded the categorical columns in the data. I selected Random Forest Regressor and XGBRegressor as my models. The data was split into two for training and testing. Both models were trained. XGBRegressor performed better than the Random Forest Regressor (This is higlighted more in the results slide). The model was deployed as Dashboard using Dash.

TOOLS USED IN THIS STEPS

1. **Jupyter Notebook**
2. **Python Programing Modules: Pandas, Seaborn, Plotly, Dash, Scikit Learn, and XGBoost**

DATA SELECTION

I used the ‘Employment by sex, rural/urban areas and economic acticity (EMP_2EMP_SEX_GEO_ECO_NB)’ indicator for analysis. From the data I used the following variable:

1. **‘Ref_label.area’.** I chose this column because it contained the country names. I renamed the column to ‘country’.
2. **‘Sex’.** I chose tis column because it higlighted the gender of the group being observed.
3. **‘Classif1.label’.** I chose this column because it contained the region description ie National, Rural or Urban. I renamed the column ‘region’.
4. **‘Classif2.label’.** I chose this column because it higlighted the different economic activies ie Total, Industry, Service and Argriculture. I renamed the column to ‘activity’.
5. **‘Year’.** This column highlighted the year the observations were made.
6. **‘obs_value’.** This column had the number of observations made in thousands.

DATA CLEANING

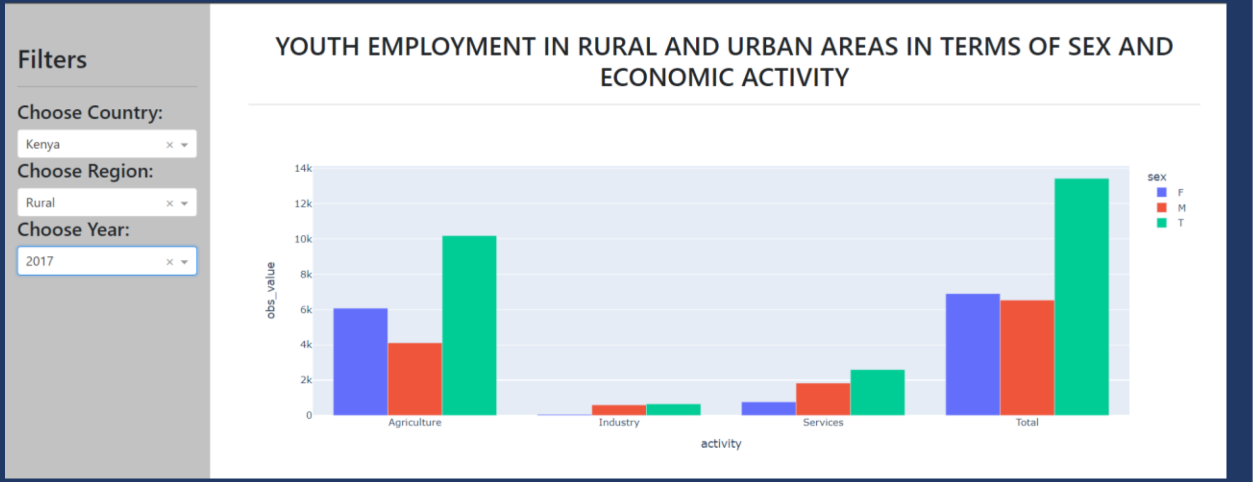
1. I dropped the columns that were not required from my analysis and remained with the highlighted columns.
2. I stripped the data in the remaining columns so has to extract the necessary data. For example in the ‘Classif1.label’ the data was represented as ‘Area Type: Rural’. This was stripped to only ‘Rural’.
3. There were no missing values in the data.

FEATURE ENGINEERING

1. I did One Hot Encoding for the categorical columns ie ‘sex’, ‘country’, ‘activity’ and ‘region’.
2. The ‘year’ column was left untouched.

Feature engineering was done using a column transformer. This is advantageous since it does all the transformations required at the same time making model deployment very easy.

DATA VISUALIZATION DASHBOARD



Filters

Choose Country: Kenya

Choose Region: Rural

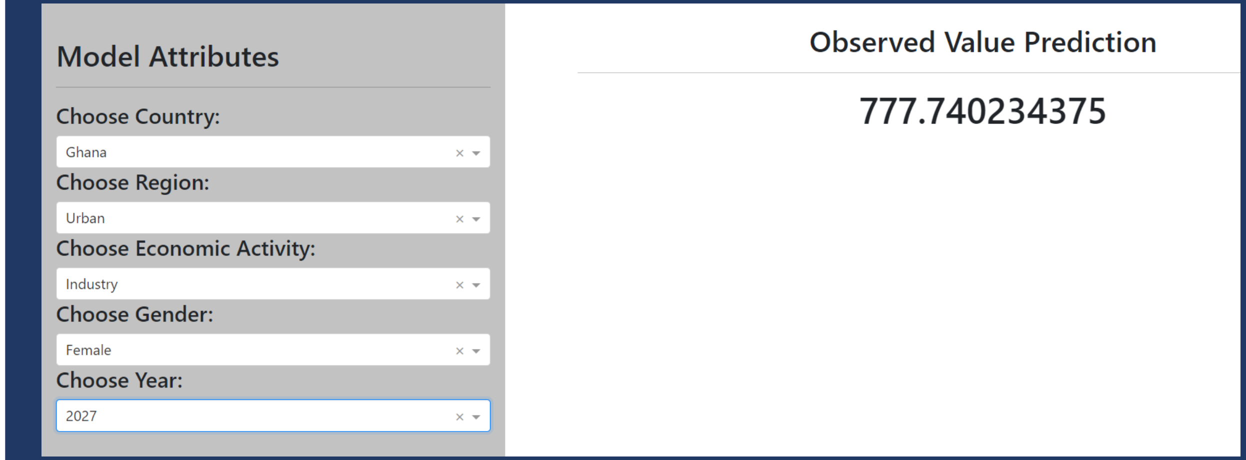
Choose Year: 2017

YOUTH EMPLOYMENT IN RURAL AND URBAN AREAS IN TERMS OF SEX AND ECONOMIC ACTIVITY

The following shows the data visualization dashboard created using Dash.

The filters are country, region and year and the dashboard display a bar graph of observed values against activity with the bars grouped by gender.

MODEL DASHBOARD



Model Attributes

Choose Country: Ghana

Choose Region: Urban

Choose Economic Activity: Industry

Choose Gender: Female

Choose Year: 2027

Observed Value Prediction

777.740234375

The following shows the model dashboard created using Dash.

The dashboard was created based on the XGBRegressor which performed better than th RandomForestRegressor ie the R Squared score for the models were:

XGBRegressor: 0.9959
RandomForestRegressor: 0.9948

The models were nealy perfect. This might be due to overfitting.

FINDINGS FROM THE DATA ANALYSIS

The following were obsereved:

1. The agricultural industry is the biggest employer of young people in the three countries. Followed by the service industry.

2. That the industrial sector is the poorest employer of young people.

3. That in urban areas the biggest employer of young people is the service industry.

4. That in rural areas the biggest employer of young people is the agricultral industry.

5. That in rural areas in Kenya and Rwanda the agricultural industry employs more females than males.

The most significant finding from the data was that there wasn't any significant difference in youth employment in all the three countries in terms of gender. The observations in terms of gender was almost equal.

RECOMMENDATIONS

From the findings some recommendations that can be made include:

1. More investments should be made in the industrial sector. This sector was observed to be the poorest employer of young people. This sector has the potential to be the biggest employer for young people as observed in other countries like India and China.

2. Also more investment should me made in the agricultural sector since it’s the biggest employer of young people in all the three countries.

POTENTIAL IMPACT OF THE MODEL

The model can be used to predict the growth of the different industrial sectors in urban, rural and nation wide. This might help policymakers in setting up various policies that might stimulate more employment of young people.