DS 3000 Final Project Presentation

SEC04 - Team #5

Tarun, Ryan, Saanika, Alex

Topic Interest and Project Research Questions

“Are there any economic indicators that could help us predict the economic status of a country?”

Our goal was to analyze data regarding the GDP Growth, Foreign Direct Investment, Unemployment Rate and Inflation Rate of different regions around the world and to use **machine learning** to see if there are any underlying indicators that drive economic growth around the world.

We'd also like to explore the idea of selecting statistics that may or may not be related in order to find out if there exists a correlation between the two. Furthermore, with information spreading about how global education levels have fallen since COVID, We want to study if it has a potential effect on the economy of the world. Our key questions are:

1. Is it possible to predict economic status given certain parameters?
2. Is there a correlation between FDI and GDP Growth? Does this differ across regions?
3. Can education levels impact economies?
4. Is Foreign Direct Investment actually beneficial for economies around the world?
5. Why are some regions of the world growing differently than others?

Our Dataset / API Call

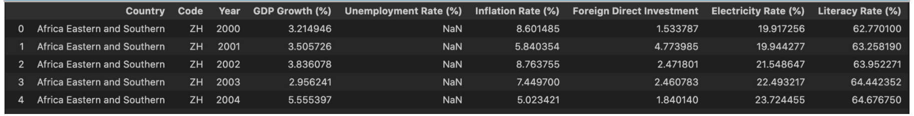
THE WORLD BANK

We used the World Bank API to gather economic data from different world regions to help understand recent trends.

URL: http://api.worldbank.org/v2/country/all/indicator/{indicator}?date={years[0]}:{years[-1]}&format=json&per_page=1000

Indicator Name	World Bank Indicator Code	Description
GDP Growth (%)	*NY.GDP.MKTP.KD.ZG*	Annual % growth of GDP
Unemployment Rate (%)	*SL.UEM.TOTL.ZS*	Unemployment rate as % of total labor force
Inflation Rate (%)	*FP.CPI.TOTL.ZG*	Consumer price inflation (annual %)
FDI Amount (% of GDP)	*BX.KLT.DINV.WD.GD.ZS*	Net inflows of foreign direct investment
Electricity Access (%)	*EG.ELC.ACCS.ZS*	% of population with access to electricity
Literacy Rate (%)	*SE.ADT.LITR.ZS*	% of people aged 15 + who can read and write

Example Data Frame of our Cleaned Data (‘Merged DF’)



	Country	Code	Year	GDP Growth (%)	Unemployment Rate (%)	Inflation Rate (%)	Foreign Direct Investment	Electricity Rate (%)	Literacy Rate (%)
0	Africa Eastern and Southern	ZH	2000	3.214946	NaN	8.601485	1.533787	19.917256	62.770100
1	Africa Eastern and Southern	ZH	2001	3.505726	NaN	5.840354	4.773985	19.944277	63.258190
2	Africa Eastern and Southern	ZH	2002	3.836078	NaN	8.763755	2.471801	21.548647	63.952271
3	Africa Eastern and Southern	ZH	2003	2.956241	NaN	7.449700	2.460783	22.493217	64.442352
4	Africa Eastern and Southern	ZH	2004	5.555397	NaN	5.023421	1.840140	23.724455	64.676750

Identifying Features in our cleaned Data Frame:

Country: Contains the region name (**e.g. Africa Eastern and Southern**)

Code: Contains the region code (**e.g. ZH**)

Year: Indicates the year the data represents (**e.g. 2000**)

Methodology of ML Models

ML Model #1 - Simple Linear Regression

For the first ML Model, we used the Linear Regression:

1. First, we set or independent and dependent variables.
   - **Foreign Direct Investment** as the X Variable (Independent var).
   - **GDP %** as the Y Variable (Dependent var).
2. Next we cross validated the dataset using **Train_test_split.**
3. Next, we **added a Bias column to the X variable** in order to calculate both the **slope and the Y-Intercept of the regression line.**
4. After adding the necessary bias column to the feature data, we used matrix multiplication using NUMPY in our '**Line_of_best_fit**' function. Using this method, we were able to calculate the Y-Intercept, and the Slope of the Regression line for 5 of the different regions we analyzed.
5. Using the "**Linreg_predict"** function, we calculated 'Ypreds', the predicted value of Y based on the regression line.
6. Using **Ypreds**, and the **actual y values**, we calculated the difference between the two to find the **residuals** for each point in the data.
7. Finally, we plotted the results, plotting the true data vs the regression line, as well as various different plots of the residuals (to determine **Normality, Homoscedasticity, and Linearity**)

ML Model #2 - Random Forest Classification

For our second machine learning model, we utilized a **Random Forest Classifier** to perform a supervised classification task. Our goal was to predict whether a country's region is classified as **"Developed"** or **"Developing"**, based on a variety of economic and social indicators we discussed above. To create the actual data for the model to train on, we classified the various regions based on research we conducted on the economic state of the regions, adding the classification to a new feature '**RegionType**'.

Before training our Random Forest ML model, we consolidated data from the five different regions in our training sample: **Africa Eastern and Southern**, **Latin America & Caribbean**, **Europe & Central Asia**, **Middle East & North Africa**, and **East Asia & Pacific**.

We started by defining our **X feature set and target variable (y).** We excluded **identifying or non-predictive features such as Country, Code, and Year, as well as Unemployment Rate (%) (due to excessive NANs in the data).**

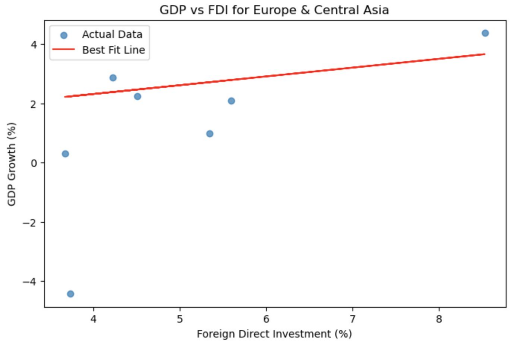
For the target variable itself (RegionType) was used as the actual data for the model, we are the foundation of the training of this model.The RegionType column, which contains categorical labels ("Developed" or "Developing"), was encoded into numerical values using from **sklearn.preprocessing import LabelEncoder**.

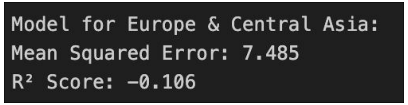
Takeaways of ML Model #1 - Linear Regression

Simple Linear Regression Model Takeaways

From the Linear Regression ML model, **we were not able to find a strong correlation** between FDI and GDP Growth %.

Example Linear Regression on Cross Validated Data





```
Model for Europe & Central Asia:
Mean Squared Error: 7.485
R² Score: −0.106
```

Challenges from Model Implementation

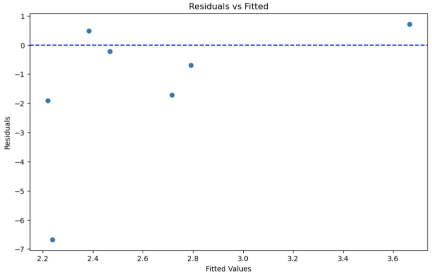
1. **Our data set was rather small since** we collected data from the years 2000 - 2023.
   a. This posed issues with the cross validation process due to the fact we used a 70% 30% split, resulting in a small training data set.

2. **Low R2 Scores**: The model result on the left indicates a very weak model fit, given the R2 score is only -0.106

3. **High MSE**: Due to the nature of the data, we believe the MSE here of ~7.5 to be quite high since the FDI and GDP Growth numbers themselves were quite small (roughly from 0 - 5).

Analysis of the Residual Plots from Model #1

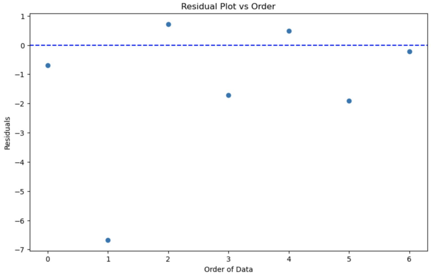
Residuals Vs. Fitted



Takeaways:

There is clear structure in the residuals, indicating a violation of the linearity assumption. The clustering suggests that the model is missing key non-linear patterns in the data.

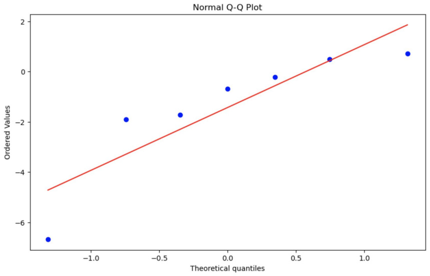
Ordered Residual Plot



Takeaways:

Subtle trends across the order of data points suggest that residuals are not independent. This could indicate temporal effects that the model does not account for.

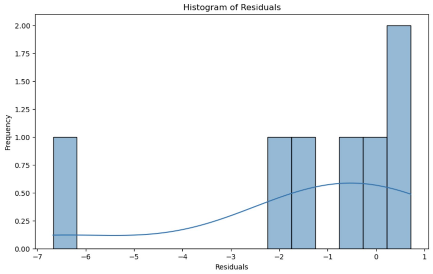
Normal QQ Plot



Takeaways:

The residuals deviate noticeably from the red reference line, which means they are not normally distributed. This violates the assumption of normality.

Histogram of Resids



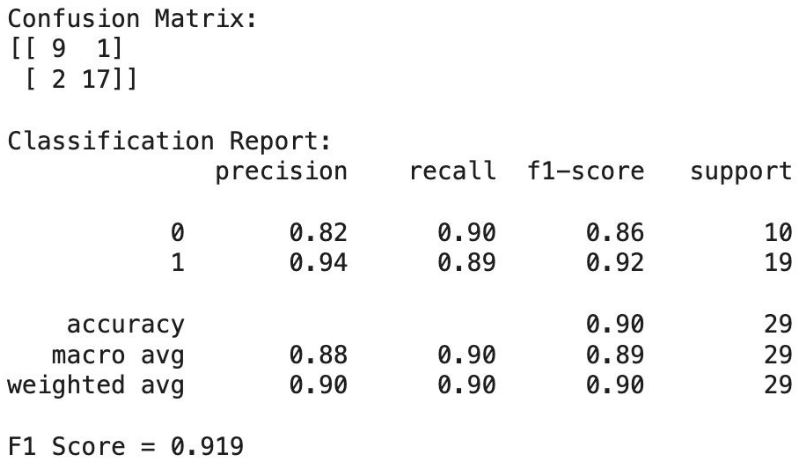
Takeaways:

The distribution of residuals is skewed and not symmetric. This further supports that the residuals are not normally distributed.

Takeaways of ML Model #2 - Random Forest

- The confusion matrix showed that out of 29 test samples, 26 were correctly classified, giving an accuracy of 90%.
- The model achieved 0.82 for precision and 0.90 for recall for the “Developing” class (encoded as 0).
- The model achieved 0.94 precision and 0.89 for the “Developed” class (encoded as 1).

Using these metrics, the model gets an **F1-score of 0.86 and 0.92 for each class respectively**, which suggests the model is well balanced for minimizing false positives and false negatives.



```
Confusion Matrix:
[[ 9  1]
 [ 2 17]]

Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.90      0.86        10
           1       0.94      0.89      0.92        19

    accuracy                           0.90        29
   macro avg       0.88      0.90      0.89        29
weighted avg       0.90      0.90      0.90        29

F1 Score = 0.919
```

The classification report containing only the “Developing” class had a perfect score across all the different metrics: precision, recall, and F1-score were all equal to 1.00. This suggests that the model was extremely effective at identifying “Developing” regions, at least in this subset. The high F1 score of 0.919 across the broader testing demonstrates the RFC’s effectiveness in handling categorical classification with economic data.

Overall Conclusions From our Project

Project Conclusions

1. **The impact of FDI on GDP growth % is still unclear on a per region basis**
   a. Due to the small sample size of years per region, it is difficult to know the true strength of the relationship
2. Even when the linear regression is run on a Data Frame that is sorted by the year and specific region, the model lacks accuracy, and fails to see a strong fit.
   a. **Even with a larger sample size ordered chronologically**, the strength of the relationship is **weak at best!**
3. Is is easy to classify countries based on the variety of socio economic data features that can be accessed on the World Bank's API.
   a. Specifically, such features as **Electricity Use %, Literacy %, Inflation %, FDI, GDP Growth %**

Potential Plans For Future Work

For future work, we would like to analyze the following:

1. **Countries on a Country by Country basis**
2. **Apply PCA to determine the most important features**
3. **Apply additional ML Models**
4. **Explore other economic data datasets**

Thank You!

SEC04 - Team #5

Tarun, Ryan, Saanika, Alex