DSN
Data Science Nigeria

2025 DSN Artificial Intelligence Bootcamp

Building Your First Predictive Model from Scratch

Predicting Youth Unemployment

Agenda

01 Understand the predictive modeling workflow (end-to-end).

02 Explore a youth unemployment dataset and relevant features.

03 Train, evaluate, and interpret a Linear Regression model.

04 Test 'what-if' scenarios using an interactive UI with policy thresholds.

Lekan Akinsande

Data, Analytics & AI Professional

lekanakinsande.com

Olalekan ("Lekan") Akinsande is a data, analytics and AI Professional. He serves as **Lead, Strategic Insights** at the Mastercard Foundation, where he **pioneers advanced analytics, predictive modeling, and geospatial solutions** to strengthen programs **tackling poverty, education, gender equity, and financial inclusion.**

Lekan's career spans consulting, research, and solution delivery across private and public sectors, including roles at **KPMG Nigeria and Data Science Nigeria**, where he led the team recognized as Africa's only finalist in the global XPRIZE Pandemic Response Challenge. Passionate about **democratizing data and AI**, he founded initiatives like *Citizen Analyst* and *withLekan* to advance data literacy and digital empowerment across Africa.

He is driven by the conviction that **data is a tool for equity, access, and agency.**

DSN Data Science Nigeria

Why Youth Unemployment Matters

- Africa has the **world's youngest population**; jobs for youth are pivotal for inclusive growth.
- Data helps **identify patterns and priority levers** for interventions.
- Predictive models **support planning and scenario analysis** (**not causal proof**).

Problem Framing & Dataset

- **Goal**: Predict Youth Unemployment Rate (%) from socio-economic indicators.
- **Features (X)**: GDP per Capita (USD), Education Index (0–1), Urban Population (%).
- **Target (y)**: Youth Unemployment Rate (%).
- **Grain (level of detail)**: country-year observations across African countries-2021.

Predictive Modeling: Core Concepts

- **Supervised learning**: learn a mapping from inputs (X) to a target (y).
- **Linear Regression**: a baseline, interpretable model for numeric targets.
- **Key idea**: fit a function that minimizes errors on training data, then test on unseen data.

Modeling Workflow (You Can Reuse This)

1. Load & Explore data (EDA).
2. Select features & target.
3. Split into Train / Test.
4. Train model (fit).
5. Evaluate (MAE, R^2).
6. Predict & Iterate (what-if).

Building the Model (in Python)

- Split: train_test_split(X, y, test_size=0.3, random_state=42).
- Model: LinearRegression().fit(X_train, y_train).
- Inspect: coefficients (direction & magnitude) and intercept.
- Interpretation: hold-other-things-constant view for each coefficient.

Evaluation Metrics (Plain English)

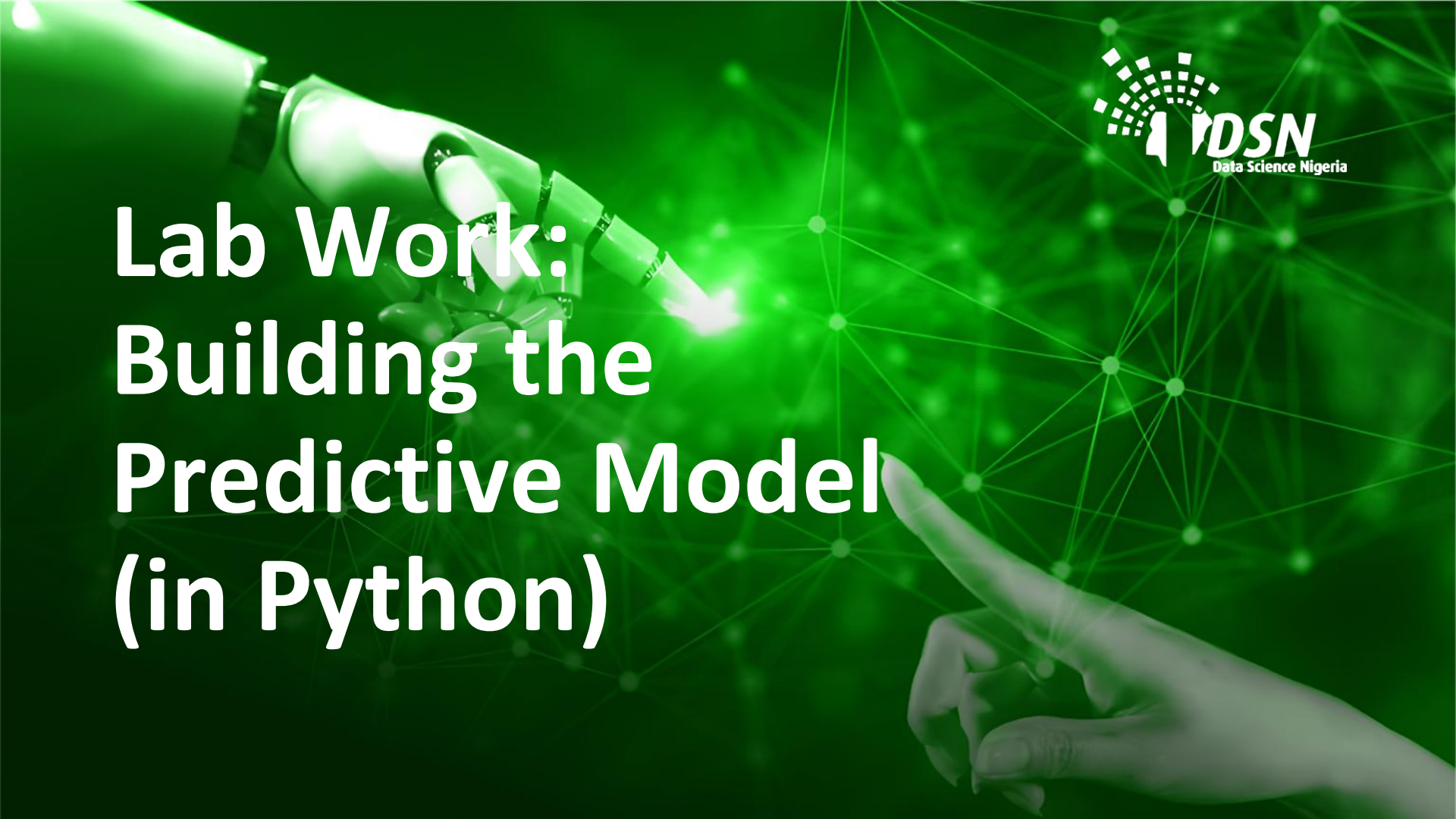
MAE (Mean Absolute Error)
– *What it is:* Average size of mistakes **in target units** (here: **percentage points**).
– *Read it like this:* “On average, we’re off by **X p.p.**.” **Lower is better.**
– *Rule of thumb:* Compare MAE to typical values (e.g., MAE 2.5 p.p. vs rates around 15–30% = **reasonable**).

R^2 (Coefficient aof Determination)
– What it is: How much of the **variation** in the target your model explains, compared to just predicting the mean.
– *Read it like this:* “Our features explain **$R^2 \times 100\%$** of differences across countries/years.”
– **Closer to 1 is better**; **<0** means worse than predicting the mean.

Visual Check: Predicted vs Actual (Scatter Plot)
– *What to see:* Points **hug the 45° line** → good fit; big spreads → underfitting/noise.
– *Patterns to watch:* Curves (try non-linear models), clusters/outliers (add features or investigate data quality).

Interactive Demo: Country → Sliders → Predict

- Pick a country to auto-fill realistic feature values.
- Adjust GDP, Education Index, Urbanization with sliders.
- Click Predict to see the rate and a policy-colored meter (0–50%).
- Reset to defaults to start a new scenario.



Lab Work: Building the Predictive Model (in Python)

Resource

GitHub - lekanakin/DSN_AI_Bootcamp25: Lecture content for my Lab Session at the DSN AI Bootcamp

github.com

github.com/lekanakin/DSN_AI_Bootcamp25

Ethics, Cautions & Good Practice

- Prediction ≠ Causation. Use causal designs for policy effect estimation.
- Data quality & representativeness matter; document assumptions and gaps.
- Avoid overfitting; validate on unseen data; consider cross-validation.
- Consider fairness and bias; check for systematic under/over-predictions.

Next Steps & Extensions

- Add richer features: internet access, access to finance, sector mix, inflation.
- Try non-linear models: Decision Trees, Random Forests; compare performance.
- Do feature importance and sensitivity analysis; communicate uncertainty.
- Publish a short brief: context, method, key insights, limitations, and next actions.

Q&A / Discussion

- Which additional features would you add first—and why?
- Where might prediction be most helpful for your work?
- How would you test if an intervention truly reduces unemployment (causal design)?

Connect with Me

Empowering Growth Through Data and Analytics Expertise | lekanAKINSANDE

lekanakinsande.com

Thank You

for being part of

2025 DSN Artificial Intelligence Bootcamp

DSN
Data Science Nigeria