DataFestAfrica Hackathon 2024: Improving Academic Outcome For Secondary Education

Overview and Objectives

In Africa, the subpar quality of elementary and secondary education has been a longstanding concern. Recent statistics reveal that approximately 76% of students who participated in the 2024 UTME scored less than 200, highlighting the need for proactive solutions. Our project aims to leverage data to improve students' performance in JAMB exams. In this project, we design and implement an enterprise data solution, predicting student performance and providing recommendations for improvement.

Our objectives are:

- Generate relevant data reflecting the African education ecosystem.
- Design an enterprise data solution for data collection, pipelining, warehousing, automation, and reporting.
- Develop an optimized model predicting student performance based on some external factors.
- Provide stakeholders with data-driven recommendations for improving student performance.

Tool:

We used Microsoft Fabric and Power BI for our end to end solution.

Dataset Selection:

To ensure a comprehensive understanding of the African education ecosystem, we sourced datasets from:

- Newspaper articles providing contextual information on education trends.
- Nigeria Stats for historical participation and facts.
- Statista
- Nigeria Budget website

Why this dataset? Since the student score we be too frictious, we decided to model our dataset to be as close to reality as possible, thus we selected and generated these datasets for their relevance, accuracy, and comprehensiveness:

- Contextual insights (newspaper articles)
- Quantitative performance metrics (JAMB data)
- Participation and failure rate analysis (Statista)
- Resource allocation context (budget and teacher-to-student ratio)

How we generated the dataset

For this project, we worked with 3 dataset:

1. UTME data from 2020 to 2024: The generated UTME scores are modeled after the official JAMB scores, reflecting the actual performance trends and statistics. The parameters used to generate the scores are based on published data from JAMB's annual reports and statistics:
    - **Failure Rate:** Ranges from 76.1% (2024) to 87.2% (2021), consistent with JAMB's reported failure rates.
    - **Above 300 Rate:** Varies between 1% (2021, 2023) and 3% (2022), reflecting the proportion of candidates scoring above 300.
    - **Score Distribution:** Scores are generated using a uniform distribution, mimicking the actual score spread and failure rate.
    - **Age and Gender Distribution:** Reflects the demographic characteristics of UTME candidates.

2. Budget 2020 to 2024: This dataset contains Nigeria's education budget allocations from 2020 to 2024, including total education budget, percentage of total budget, and breakdowns by Federal Ministry of Education, UBEC, and TETFunds. This is gotten from the budget reports.

3. The generated Teacher-Student Ratios are modeled after realistic ranges, with public schools (20:1 to 100:1) and private schools (15:1 to 50:1), reflecting variations in resource allocation across Nigerian states.

By using published and official data to inform the our data generation process, the resulting dataset provides a realistic representation of UTME performance trends, enabling meaningful analysis and insights.

Why did we go this way?

We prioritized Resource Allocation, budget, and Historical performance evaluation over generating mock/test scores and demographic data because randomly generating student scores and demographic data would serve no practical purpose, lacking real-world relevance and validity.

However, analyzing resource allocation and historical performance provides actionable insights into systemic challenges. This approach is also backed by educational research emphasizing the impact of resource allocation on student outcomes (Hanushek, 2003; OECD, 2012). We figured that focusing on these factors, our model offers data-driven recommendations for meaningful improvement.

Machine Learning:

Objective: The notebook is a machine learning solution aimed at addressing the problem of student failing exams in Nigeria. The primary objective is to develop an accurate model that can determine the probability of a student passing or failing an exam based on features such as budget allocated by the FG to the education sector, age, sex amongst others. The notebook includes steps for data extraction, transformation, modeling, and inference.

ETL Process:

1. Extract: The datasets (comprising a merge of the "nigeria_education_data_2020_to_2024" and the "budgetedu" dataset) used to train the model were pulled from the Lakehouse in Microsoft Fabric. The data is extracted using PySpark's SQL endpoint and converted to the standard Python libraries (pandas) for further data pre-processing.
2. Transform: Feature selection is performed based on correlation, the 'Score' column although had the highest correlation was dropped because based on domain knowledge if the Score is known beforehand, it defeats the purpose (in other words it is also the target column). Log1 transformation was employed on the dataset to curb the imbalance in features.
3. Load: Transformed data is loaded into the machine learning models for training. Data is split into training and testing sets using an 80-20 split.

Data Modeling:

1. Model Type: The chosen model is the Logistic Regression.
2. Feature Selection: Feature selection is done based on correlation analysis and importance ranking.
3. Model Training: Asides Logistic Regression, other classification models such as XGBoost were employed giving an accuracy score of approximately 0.41448. The Logistic Regression Model first experiment gave an accuracy score of approximately 0.7918, but the recall, precision and F1 score of the 1s class was 0.00 which pointed to the fact that it was doing a bad job at predicting the probability of the 1 class, so we applied SMOTE to curb the imbalance further.
4. Evaluation Metrics: After the application of SMOTE, Accuracy score dropped to approximately 0.4732, a testament to the fact that the features are doing a bad job at correlating with the target variable and are not enough to achieve our ML objective. But interestingly the recall, precision and F1 score of the 1s class improved to 0.66, 0.23, 0.34 respectively meaning the SMOTE application helped curb the imbalance.

Inference:

1. Deployment: The trained model is used for inference on new patient data. Inference is performed by loading the model, passing in new data, and interpreting the output as either positive (presence of heart disease) or negative (no heart disease).
2. Model Updates: Retraining strategy includes updating the hyper-parameters and re-validating the model.
3. Run Time:

    Data Preprocessing: approx. 2 minutes.
    Model Training: Logistic Regression: approx. 2 minutes, Logistic Regression with SMOTE: approx. 3 minutes.
    Inference: < 1 second per new data point.

Limitation: The features on the dataset have very poor correlation values (<0.1) with the target variable, therefore there is a need to gather features that directly affect the target before a very good classification model that can directly predict students passing or failing can be deployed.

Performance Metric:
Logistic Regression Model:

Accuracy score: 0.4732
Precision, Recall and F1 score of the 0 class respectively: 0.83, 0.42, 0.56
Precision, Recall and F1 score of the 1 class respectively: 0.23, 0.66, 0.34

Data Visualization:

Microsoft PowerBI was employed to create insightful visuals.

Data Source methodology: Data was connected using a direct query from the warehousing in Microsoft Fabris utilizing the SQL Server endpoint.

Visuals:
Metrics measured:

1. **Annual Performance by Top State:** Analyzed student performance in top-performing states.
2. **Failure Rate:** Calculated the percentage of students who did not meet performance standards.
3. **Success Rate:** Determined the percentage of students who achieved desired performance levels.
4. **Annual Average Score by Gender:** Compared average scores between male and female students across different years.
5. **Annual Rate:** Evaluated overall student performance trends over time.