Optimised Selection Process For SAND Learning Programmes

Prepared by Otaku Oracle (Team 9)

Explore AI Academy

Sand Tech Internship Project

White Paper

2024

Table of Contents

Project Overview

- **Purpose**: The Optimised Selection Process for Learning Programmes project aims to improve learner success rates across SAND’s programs by leveraging data-driven insights. This project will refine the selection process, enhance learner completion rates, and identify best-fit candidates, ultimately improving learning outcomes and participant experiences.
- **Problem Statement**: SAND’s current selection process underutilizes application data, leading to potential onboarding of learners who may struggle to complete programs. Additionally, decentralized learner data complicates reporting, analysis, and timely insight generation, resulting in higher dropout rates and lower completion rates, impacting overall program effectiveness.

Value to SAND

By optimizing learner selection and data systems, this project will:

- **Boost Completion Rates**: Enhance intake quality to admit highly qualified, committed learners, increasing program completion.
- **Reduce Dropouts**: Identify early risk factors to improve retention through a refined intake process.
- **Elevate Program Quality**: Admit candidates more likely to excel, fostering a satisfying and successful learning environment.
- **Drive Continuous Improvement**: Set a precedent for operational excellence and data-driven selection within the industry.

Through these improvements, SAND can better support learners, enhance operational efficiency, and amplify the quality and impact of its educational programs.

Proposed Solutions Summary: Optimizing SAND’s Learner Selection and Data Management

This project introduces a multi-dimensional approach to transform SAND’s learner selection and data handling processes, focusing on enhancing selection accuracy, improving data accessibility, and elevating program success rates. The core objectives are:

1. **Applicant Ranking System**: Leveraging data collected at the application stage, this ranking system prioritizes applicants based on their likelihood of program success. This streamlines the intake process and increases learner outcomes by admitting the most promising candidates.
2. **Centralized Learner Data**: Establishing a centralized data repository enables efficient reporting and analysis. This allows SAND to track learner progress, monitor program performance, and assess success rates through accessible metrics and insights.
3. **Dashboard for Reporting and Insights**: An interactive dashboard visualizes data on learner performance, dropout trends, and overall program efficiency. This tool supports the executive committee with real-time data for informed decision-making and fosters ongoing strategic improvements.
4. **Executive Committee Reports**: Regularly produced, these data-rich reports offer concise, actionable insights into learner outcomes and program efficacy, empowering the executive team to make evidence-based adjustments in both learning programs and selection strategies.

Industry-Leading Talent Identification and Acquisition

The ultimate vision is to establish an industry-standard talent identification system, utilizing advanced data analytics and machine learning to improve applicant selection. This not only positions SAND as a leader in efficient learner on-boarding but also in setting new benchmarks for selection processes.

Data and Methodology

AICE Datasets

- Assignment DataFrame: 9 columns, 407333 rows, size 110430082 bytes
- Learning DataFrame: 21 columns, 150535 rows, size 130058784 bytes
- Applicants DataFrame: 33 columns, 354152 rows, size 40295479 bytes

Build an Application Ranking System

Data: Gather historical data on applicants and existing learners (demographics, prior academic performance, application data, etc.).
Model: Use machine learning algorithms to rank applications based on a set of predefined criteria. Algorithms like Random Forests or Gradient Boosting could be effective here, using both categorical and numerical data.
Outputs: A ranked list of candidates based on predicted success in the program.

Dropout Prediction Models

Data: Use learner performance data (attendance, assignment submissions, grades) to predict the likelihood of dropout.
Model: Classification models such as logistic regression, decision trees, or neural networks can be used for dropout prediction.
Outputs: A probability score indicating the likelihood of a learner dropping out.

Workflow Automation with Airflow

Set up **Airflow** to manage the data pipeline and automate processes such as:

- Data extraction from **S3**.
- Training and retraining of machine learning models (application ranking, dropout prediction).
- Monitoring learner performance metrics and sending alerts for early interventions.

Deploying on AWS EC2 and S3

Use **S3** for:

- Uploading and storing datasets.
- Storing the results of the models (ranked applications, dropout predictions).

Set up an **EC2** instance for:

- Hosting the machine learning models and dashboards.
- Mounting the **S3** bucket to access datasets in real-time.
- Running Airflow and Docker containers for task orchestration and isolated environments.

Database Design Using SQL

- **ERD (Entity-Relationship Diagram)**: Designing an ERD, especially for working with a structured relational database.
- **SQL Database**: Store the ranked applicant data and learner performance data, which will help generate insights via queries.

Build a Dashboard for Monitoring

Why build a dashboard?

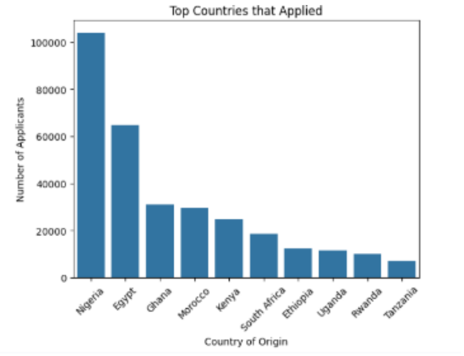
- Displaying insights from the application ranking system.
- Monitoring learner progress and dropout prediction scores.
- Providing a visual tool for stakeholders to track overall program effectiveness.

What to include:

- Visualize **application rankings**, showing top applicants based on predictive models.
- **Dropout risk**: A chart showing learners most at risk.
- **Real-time monitoring**: Integrate with **Airflow** to show pipeline status (data updates, model retraining).
- **Progress tracking**: Display overall progress of learners and program outcomes.

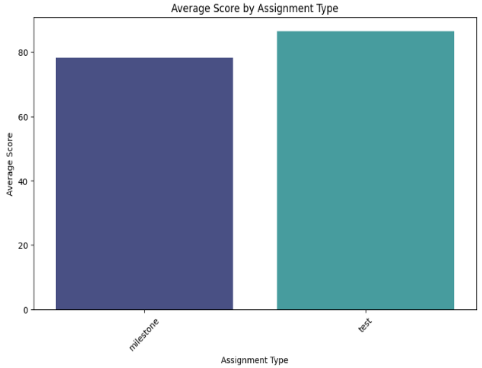
Results and Key Insights

Data Exploration Summary



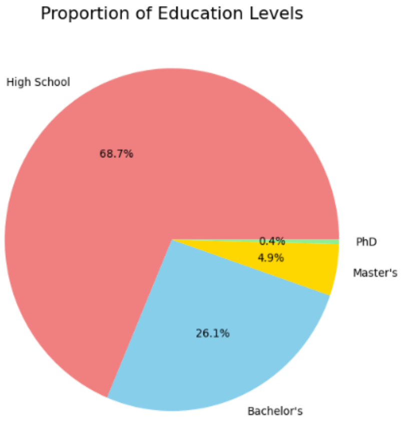
Applicants Demographics

Analyzing applications by country helps SAND understand geographical diversity and address region-specific needs. This allows SAND to identify application trends, access barriers, and preparedness levels, tailoring support for greater accessibility and completion rates. Adapting resources for diverse backgrounds ensures programs remain inclusive and impactful across regions.



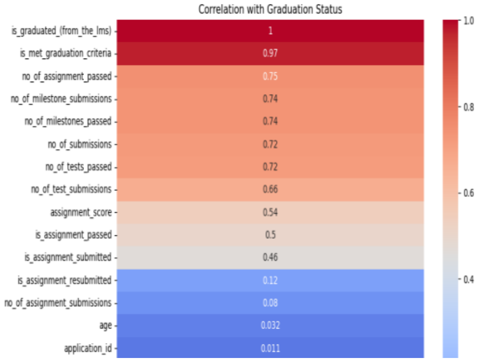
Assignment Scores

Analyzing assignment scores in relation to the 65% pass mark is essential for evaluating learner performance and program effectiveness at SAND. It helps identify trends in student achievement, highlights areas of difficulty, and assesses the rigor of assignments. This analysis informs targeted interventions for at-risk students, ultimately enhancing overall learner outcomes and completion rates.



Academic Backgrounds

Analyzing applicants' academic backgrounds helps SAND create targeted engagement strategies that resonate with each learner's experience level. By offering additional support for those with less formal education and more advanced materials for highly educated learners, SAND can foster a more engaging and inclusive environment. This approach keeps learners actively involved, promotes a sense of belonging, and strengthens motivation, ultimately improving both retention and completion rates.



Correlation Analysis

By calculating correlation coefficients, SAND can identify which factors are most strongly associated with learner success and completion rates. This analysis aids in feature selection, ensuring that the most relevant variables are used in predictive models for the applicant ranking system. Additionally, visualizing these correlations can uncover insights into how various applicant characteristics interact, guiding targeted interventions and support strategies to enhance learner outcomes.

Future Work and Improvements

Model Building with DNN and NLP

- Deep Neural Network, Input Layer: Include both structured numerical data (like grades, and demographics) and the feature vectors from the textual data
- Hidden Layers: Use multiple dense layers to capture interactions between different features. Output Layer: This could be binary (for predicting whether a student will complete a course) or regression (for predicting completion rates or grades)

NLP Pipeline

- Apply NLP models to analyse qualitative feedback from students to uncover patterns like satisfaction with courses or specific issues impacting performance
- Analysis if Key Feature: Use techniques like feature importance analysis to understand the most influential factors for student success
- Demographic Influence: Analyse how different regions or qualifications affect performance
- Course-Based Patterns: Identify courses that are consistently linked to higher or lower success rates

Performance Metrics

- **Metric Creation:** Track completion rates for each cohort to analyze whether the selection system improves these rates over time. The accuracy of the applicant ranking system will be evaluated by the correlation between expected applicant ranks and their actual performance in the program, reflecting the effectiveness of the selection process in choosing successful candidates.
- **Metric Review:** Cross-check key performance indicators against real data to assess accuracy. If the ranking algorithm is overly simplistic, it may rely on surface-level criteria (e.g., region, prior qualification) that do not fully predict success.
- **Metric-Loss Comparison:** Evaluate Applicant Ranking Accuracy by using a classification loss function, such as cross-entropy loss, to measure the difference between predicted success likelihood and actual outcomes. A low loss indicates that predicted rankings closely align with real student performance.
- **Custom Metric:** Implement a rank-based loss (e.g., pairwise ranking loss) to prioritize candidates based on correct ordering, ensuring the ranking system accurately reflects real-world outcomes, where top-ranked students perform better.

Conclusion

In conclusion, the Optimised Selection Process for Learning Programmes project provides a comprehensive framework for addressing the challenges faced by SAND in selecting and supporting learners. By implementing a data-driven applicant ranking system and centralizing learner data, SAND can significantly improve completion rates and overall program effectiveness. We encourage stakeholders to embrace these insights and invest in the proposed solutions, fostering a culture of continuous improvement. As SAND pioneers innovative practices in learner selection, it will not only enhance educational outcomes but also set a benchmark for excellence in the field. Together, we can transform the educational landscape and ensure that every learner reaches their full potential.

References

- *Optimised Selection Process For Our Learning Programmes* Project Documentation By *Kwazi Mthembu* & *Mark Mutaiti (2024)*

Contributors

❖ Sandiso Magwaza – Data Scientist (Team Lead)
❖ Lulama Nelson Mulaudzi- Data Scientist (Project Manager)
❖ Micheal Thema- Data Scientist (Github Manager)
❖ Hlamulo Ndlovu - Data Engineer
❖ Sinawo Londa - Data Scientist
❖ Josephina Ndukwane - Data Scientist
❖ Nolwazi Vezi - Data Scientist
❖ Phumzile Nomthandazo - Data Scientist