SESUG 2023 PAPER 187

Empowering Decision-Making: Predictive Analytics Integration with SAS® and Python™ in a Higher Education Setup

Vindhya Hegde, MS, MBA, NC State University

ABSTRACT

The rapid advancement of data-driven technologies has revolutionized the way higher education institutions analyze and interpret complex datasets. This paper presents a complete approach that combines SAS® Data Integration Studio (DI Studio), SAS® Visual Analytics (VA), and Python™ for modeling, data visualization, and predictive analytics in the context of data elements in higher-ed.

SAS® DI Studio serves as the primary tool for merging and transforming diverse data sources into organized datasets. It provides a robust set of features to facilitate data preparation. The transformed data is then visualized using SAS® VA, which provides interactive visualizations tools. By using various visualization tools like graphs and charts, we can analyze the data and discover patterns. This gives us the ability to gain insights into historical patterns, which can be further used to predict the future.

For predictive analytics, Python™ is used as a primary tool. Utilizing libraries such as scikit-learn, pandas, seaborn, and matplotlib in Python™ enables the development and application of predictive models. Leveraging machine learning algorithms, these models utilize the historical data to forecast trends, enabling stakeholders to make informed decisions and develop strategies for the next three years.

By integrating SAS® DI Studio, SAS® VA, and Python™, this approach provides a comprehensive solution for utilizing historical data to forecast future trends. The findings presented in this paper demonstrate the value of leveraging these powerful tools in higher education, enabling institutions to anticipate and plan for the future, optimize resource allocation, and make data-driven decisions.

INTRODUCTION

Data driven technologies have brought about a change in how educational institutions evaluate and analyze large scale datasets. This paper presents a comprehensive approach for data modeling, data visualization, and predictive analytics within the context of higher education, leveraging the combined power of SAS® Data Integration Studio, SAS® Visual Analytics, and Python™.

The decision-making processes used by educational institutions are driven by student data, which provides crucial information about applications, enrollment patterns, and academic achievement. In addition to gathering and organizing this data, the technique presented in this paper shows how to take advantage of its potential to guide proactive decision making.

Our approach has three phases that all work together to achieve the overall objective of data driven excellence in higher education:

- DATA MODELING USING SAS® DATA INTEGRATION (DI) STUDIO
- DATA VISUALIZATION USING SAS® VISUAL ANALYTICS (VA)
- PREDICTIVE MODELING USING PYTHON™

The objective of this project is to provide data-driven insights and recommendations that can help NC State University to enhance its recruitment and enrollment strategies for international students, and to improve its ability to attract and retain a diverse and talented international student population.

DATA MODELING USING SAS® DI STUDIO

The Data Modeling stage further includes these steps:

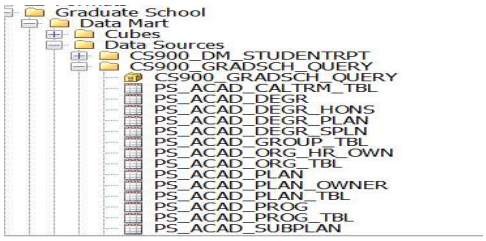
1. Data Collection
2. Data Manipulation
3. Data Upload

DATA COLLECTION

The business requirement for this project was to pinpoint nations that the Department of Outreach and Diversity may strategically pursue for potential enrollment of international students. The project began with a thorough data collection strategy that laid foundation for effective decision making.

Two diverse but complementary sources, each of which contributed crucial insights to project's success, served as the foundation of the data collection operations.

- **Admissions Table:** This oracle table consisted of data from past ten years on applications, admissions, and matriculations of international students. In addition to giving the analytical process the crucial historical context required for predictive modeling and strategic planning, it gave a comprehensive view of enrollment trends across time.
- **Demographics Table:** The Demographics table, which is a part of the Oracle database, also consisted of past ten years data, was a key table in this project. It provided crucial data of international students, covering information on things like citizenship country, race, gender, and more. Understanding the diversity and makeup of the international student body, a key component of targeted recruitment efforts, was made possible with this demographic information.
- **Data Curation and Integration:** Creating Comprehensive Dataset: Using these two rich data sources, the datasets underwent careful curation inside SAS® DI Studio. Here, the datasets were selectively picked and tailored to the specific requirements. The data curation process comprised of data cleansing, modification, and manipulation. The goal was to make sure the data was consistent, trustworthy, and ready for further analysis. Additionally, the chosen datasets were merged with the original information to create a comprehensive repository that captured the whole picture of international students' admissions and demographics. This standardized dataset served as the foundation for the analytical work.



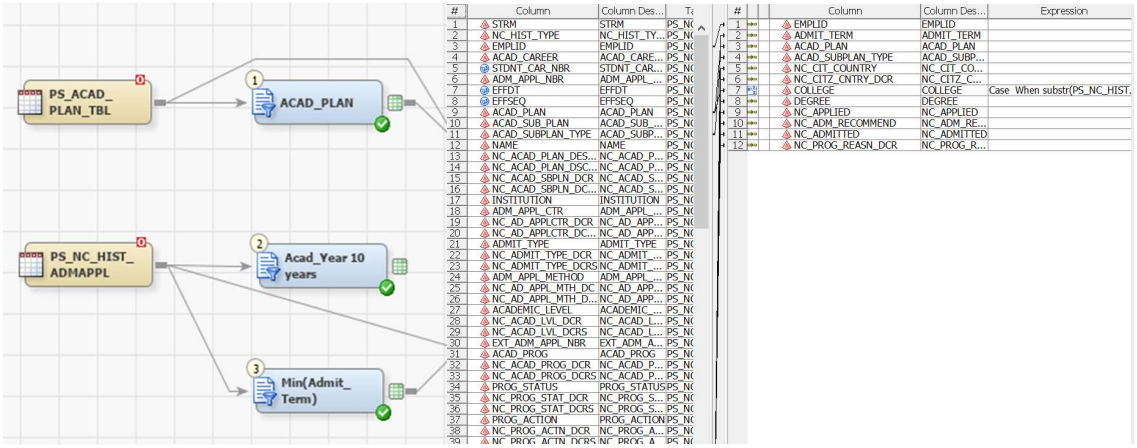
Display 1: Screenshot of Data Collection from data sources folder

DATA MANIPULATION

Data Manipulation is a crucial step in our data analysis process. During this stage, the raw data is refined and organized using a variety of techniques and methodologies in order to meet the business requirements.

Extraction of Fields:

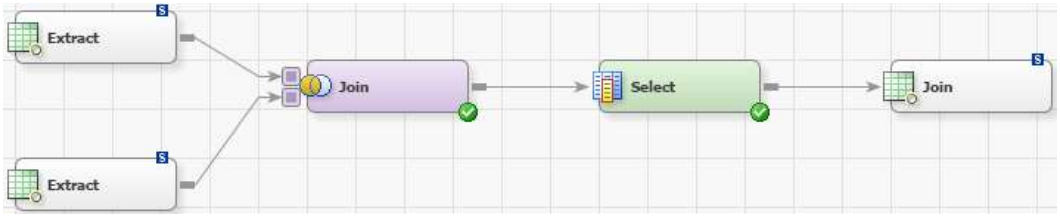
There were large number of fields, ranging in number from 100 to 150, in each of the two main tables, Admissions and Demographics. Only the fields that were relevant to our goals were carefully isolated, in order to streamline the data and make it more conducive to our study. Additionally, we used transformations as appropriate to guarantee that the gathered data complied with the precise business needs.



Display 2: Screenshot of extraction of fields from two tables

Combining Two Fields:

Within the SAS® environment, the next crucial step included the merging of the Admissions and Academic Plans' data. A left outer join technique was used to do this, producing a consolidated dataset. To create the final dataset that would form the foundation of our study, the fields within the resulting dataset were rigorously mapped.



Display 3: Screenshot of combining two tables using left outer join

Field of Study:

Managing a large number of academic programs can be a daunting task. To simplify this process, we employed a grouping mechanism based on the common Classification of Instructional Programs (CIP) codes found within the database. This allowed us to categorize the academic programs effectively, reducing complexity and facilitating a more streamlined analysis. A code snippet illustrating this grouping process can be observed in the provided figure.

```
CASE

When SUBSTR(CIP_CODE,1,2) IN ('16','23','24','38','50','54') then 'Arts and Humanities'
When SUBSTR(CIP_CODE,1,2) IN ('54','01','03','26','31') then 'Biological and Agricultural
Sciences'
When SUBSTR(CIP_CODE,1,2) IN ('52') then 'Business'

When SUBSTR(CIP_CODE,1,2) IN ('13','31','42') then 'Education'
When SUBSTR(CIP_CODE,1,7) IN ('51.0913','51.2309') then 'Education'

When SUBSTR(CIP_CODE,1,2) IN ('14','15') then 'Engineering'
When SUBSTR(CIP_CODE,1,5) IN ('40.10') then 'Engineering'

END as Field of Study
```

CIP_CODE	FIELD_OF_STUDY
11.0802	Mathematics and Computer Sciences
00.0000	Unknown
30.2001	Other Fields
30.2001	Other Fields
30.2001	Other Fields
30.2001	Other Fields
30.2001	Other Fields
30.2001	Other Fields
30.2001	Other Fields
30.2001	Other Fields
30.2001	Other Fields
30.2001	Other Fields
32.0109	Other Fields
32.0109	Other Fields
30.2001	Other Fields
30.2001	Other Fields
00.0000	Unknown
00.0000	Unknown
11.0102	Mathematics and Computer Sciences
45.0201	Social and Behavioral Sciences
14.0501	Engineering
52.0201	Business
26.1201	Biological and Agricultural Sciences
14.0901	Engineering
26.1301	Biological and Agricultural Sciences
03.9999	Biological and Agricultural Sciences
30.3001	Other Fields
01.1001	Biological and Agricultural Sciences

Display 4: The manipulation of Field of Study using CASE statement and screenshot of the final field

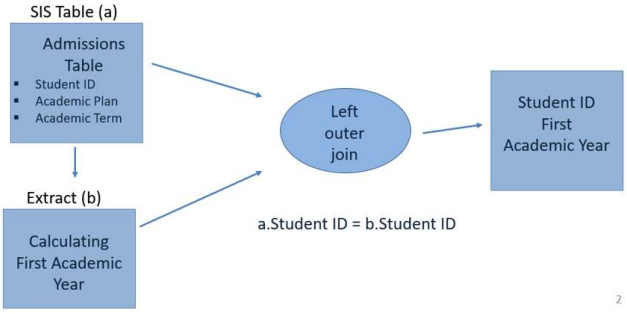
Dynamic Calculation of First Term: A Decade-Adaptive Formula

In order to do this, we created a dynamic algorithm that accurately determines the first term of enrollment and smoothly adjusts with the passing of each decade. This formula assures that the calculation is correct and current regardless of the current year by including the CATS function and utilizing SAS®®'s adaptability.

To find the students' initial academic year, this formula first takes the year proportion of the maximum admission term ('ADMIT_TERM') and subtracts 10 years. To properly identify the first term, the formula additionally extracts information about the semester. As an illustration, the formula dynamically computes the first term as 2013, aligning with the data from the previous decade for analysis, when used in the year 2023. Similarly, it easily adapts to provide a first term of 2014 in 2023.

Formula :

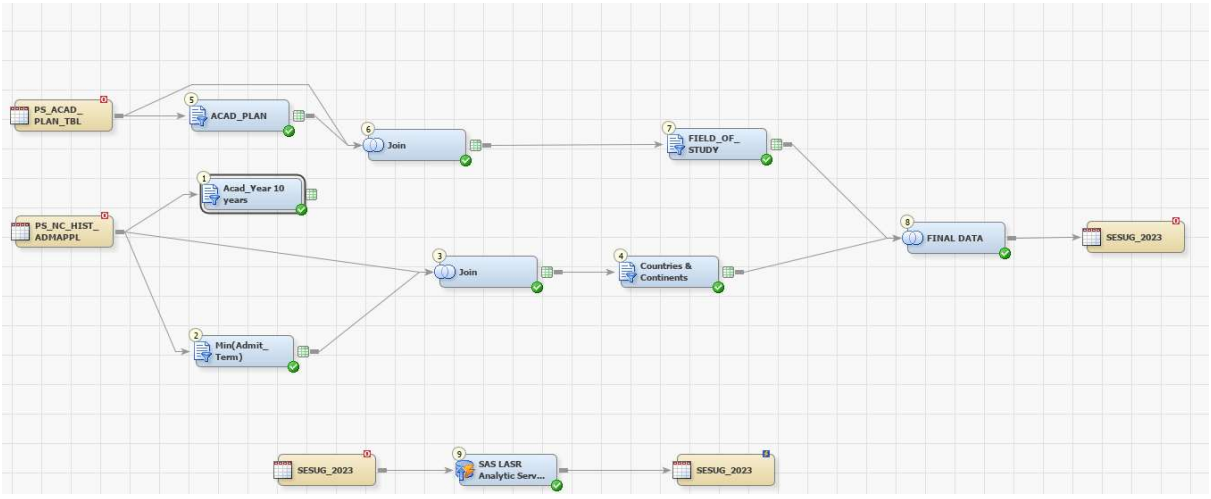
```
CATS('2',PUT(INPUT(SUBSTR(max(ADMIT_TERM),2,2),4.)-10,4.),SUBSTR(MAX(ADMIT_TERM),4,1))
```



Display 5: Formula for calculating the first academic year and figure explaining the final extract

Full Dataset Modeling

After data manipulation, the modeling of the dataset was done to obtain the final dataset.



Display 6: Screenshot of full data modeling

DATA UPLOAD

The final step was to upload this data to the SAS® LASR Server. The registered table that was created in the previous step was then pushed into the SAS® LASR Server environment using the LASR upload transformation under the SAS® Data Integration Studio.

Display 7: Screenshot of how the data was uploaded using SAS® LASR

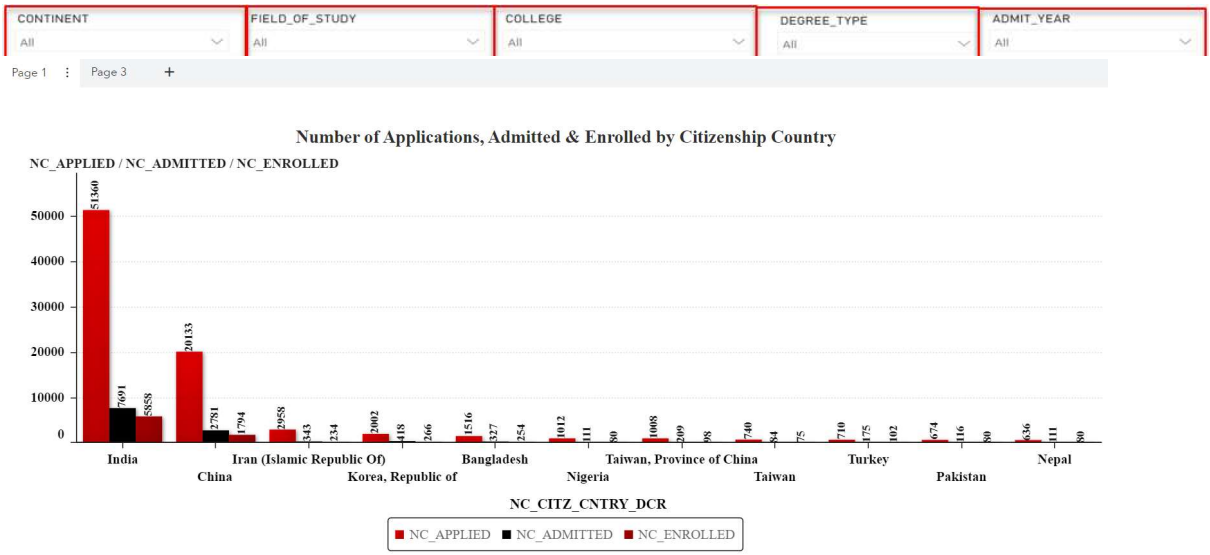
DATA VISUALIZATION USING SAS® VA

In this section, we examine how SAS® Visual Analytics (SAS® VA) dynamically transforms raw data into useful insights. As our journey progresses, SAS® DI Studio, where data is cleaned up and prepared before entering SAS® VA, is seamlessly integrated.

Creating a Comprehensive Dashboard

Creating an interactive dashboard was one of the project's key factors. This dashboard acts as a primary core of insights, making complicated data quickly understandable. The rigorously prepared and loaded data is then visualized by applications, recommendation, and enrollments for international students.

The top 10 countries according to applications, recommendation, and enrollments indicators are prominently shown on the dashboard. With the help of this information, stakeholders can pinpoint the geographical areas that have the highest levels of engagement, laying the groundwork for more focused recruitment campaigns.



Display 8: Screenshot of Dashboard in SAS® VA with top 10 countries and slicers on the top

We added a dynamic slicer component to increase user interaction and enable data exploration from various perspectives. By selecting from a dropdown menu that includes a variety of criteria, this slicer enables users to distinguish between different types of data. By Continents, Degree Type, Field of Study, and College, users can slice the data. Due to this versatility, stakeholders can customize their studies to certain criteria, resulting in more targeted and data-driven decision-making.

Enrollment Trends:

Our second dashboard page dives into past trends beyond the current snapshot. Here, trend lines take center stage, showing the trajectory of applications, recommendations, and enrollments for international students over time. Users can filter and examine these trends depending on their individual areas of interest using the same dynamic slicer that was presented on the previous page.

These trendlines serve as a useful tool for forecasting future patterns in addition to revealing past performance. A proactive recruiting strategy and resource allocation are made possible by the interaction of historical data and predictive analytics, which enables stakeholders to foresee changes in application, recommendation, and enrollment numbers.

Display 9: Screenshot of Dashboard in SAS® VA with trends and slicers on the top

SAS® VA has proven to be a crucial tool in our search for data-driven insights in higher education. Our use of SAS® VA demonstrates its ability to convert complicated data into useful knowledge, eventually influencing our institution's future recruitment and enrollment of overseas students.

PREDICTIVE MODELING USING PYTHON™

Predictive modeling offers institutions the ability to foresee and plan for the future in the context of higher education. This part examines our attempt at predictive modeling, which is based on historical data of the past ten years and makes use of linear regression machine learning paradigm. Our attempt is aimed at understanding of historical trends and projection for the next three years of application, recommendation, and enrollment of international students.

Analyzing a Decade of Historical Data:

The first step in predictive modeling is to delve into the historical data set that covers the last ten years. This information captures the sensitivities, variations, and trends that have influenced the student enrollments and recruitment strategies.

Machine Learning Concept: Linear Regression:

The machine learning idea of linear regression is at the core of our predictive modeling. This method was chosen because of its potency in scenarios where there is a relationship between one or more independent factors, such as year, and a dependent variable, in this case, the number of applications, recommendation, and enrollment. Given that this relationship may be roughly represented by a straight line, linear regression is an appropriate method for forecasting using past data.

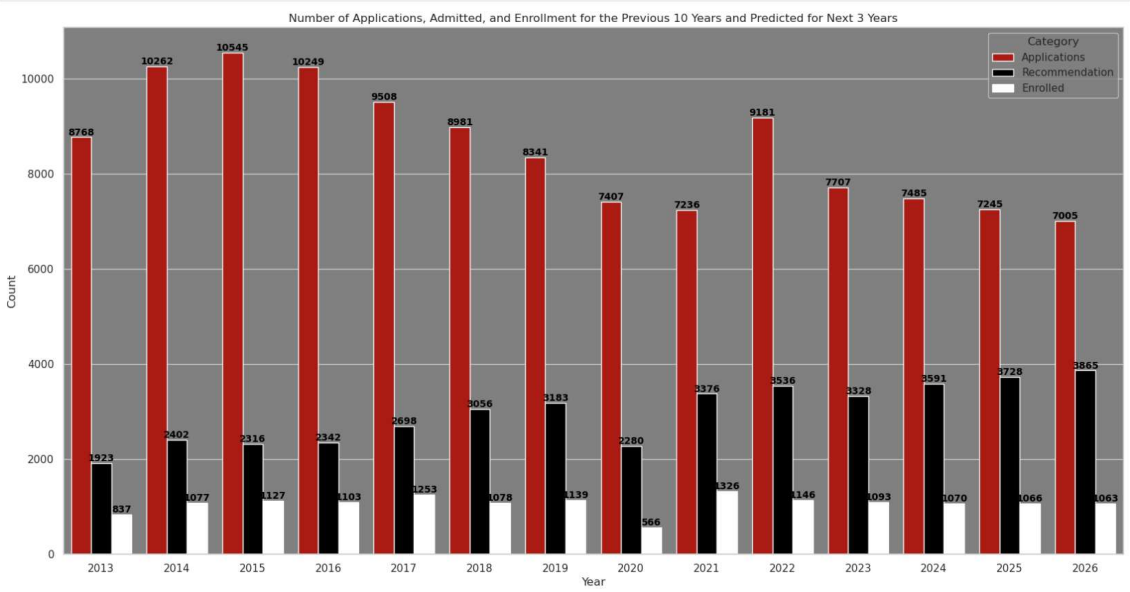
Uncovering Insights for the Next Three Years:

We start the process of predicting the future with our Linear Regression. In order to comprehend how the number of applications, recommendation, and enrollment from

international students has varied over the past 10 years, this model learns from historical data. It finds trends, recognizes patterns, and calculates the influence of time.

Empowering Proactive Decision-Making:

We are able to take proactive action with this predictive ability. It offers evidence-based insights into what the future may contain, going beyond speculation. Resources can be wisely allocated, recruitment tactics can be adjusted, and decision makers can get ready for shifting dynamics in trends of international students.



Display 10: Graph of previous 10 years and the predicted next 3 years using Python™

CONCLUSION

In conclusion, we enable educational institutions to make data-driven decisions with comprehensive outcomes by carefully gathering, curating, and visualizing data. These tools are vital for maximizing resource allocation and recruitment tactics since they not only give insightful analyses of past trends but also the ability to predict future trends. This also influences the effects of how international students will be enrolled in higher education institution, how resources will be managed, and how effective it would be for both students and universities.

Furthermore, the proper usage of these technologies also highlights how data-driven approaches in higher education have potential to excel. It gives institutions the ability to respond to resources allocations, changing environment, and overall improvement. This strategy serves as a testament to the profound impact that data driven decision making can have on determining the future of academic institutions and their capacity to flourish in an increasingly competitive global landscape, we navigate the complex challenges and opportunities in higher education.

CONTACT INFORMATION

Vindhya Hegde – Data Analyst/BI Analyst – Information Management
E: vhegde2@ncsu.edu