Charting Pathways: Predicting Preliminary Year Pass Rates at UMass Dartmouth

Authors: Veda Sahaja Bandi, Sindhuja Baikadi

The Issues:

In this analysis, our focus centers on understanding what helps students succeed in moving from a preliminary year to their chosen degree program at UMass Dartmouth. This dataset, while comprehensive, presents challenges including missing values and the selection of relevant predictor variables. Our goal is to identify key factors that predict a student's likelihood of successfully transitioning from the preliminary year to their chosen degree program. As we delve into the dataset using logistic regression, several questions arise:

- How do we handle missing data in crucial variables that are fundamental for predicting student success?
- Determining the optimal set of factors that are most important for predicting a student’s success in passing the preliminary year.
- Can our analysis offer meaningful insights into the factors influencing student success in the preliminary year?

Addressing these inquiries is crucial for identifying predictive trends in the data and establishing a foundation for further analysis into the dynamics of student success in preliminary-year programs.

Findings:

The study of students passing the preliminary year at UMass Dartmouth revealed important factors influencing their success in transitioning to their chosen degree program. Despite data challenges, such as missing information, we identified key predictors associated with student success. The missing data was removed as it did not impact the results hugely.

Variables related to student engagement, academic readiness, and socio-demographic background like the Number of Workshops Attended, whether the student completed Campus Event Requirements, etc, emerged as significant indicators of success. This suggests that both intrinsic factors (e.g., academic readiness) and extrinsic factors (e.g., support services accessibility) are influential.

Our analysis offers valuable insights into the factors influencing student success in the preliminary year. Understanding the implications of these factors is essential for developing tailored interventions to improve student success rates.

These findings underscore the importance of targeted support and interventions for students, particularly those identified as at risk based on the predictive model. By understanding the variables that significantly impact student success, UMass Dartmouth can tailor its resources and programs more effectively.

Discussion:

The study reveals patterns in student success in the preliminary year, suggesting that a multifaceted approach is necessary to support student transition into degree programs. Success is influenced by a combination of academic, personal, and socio-economic factors, which raises critical questions about the support structures in place for students during this transitional phase.

The pronounced impact of certain predictors indicates potential areas for intervention, such as enhanced academic advising, personalized support services, and socio-economic assistance. Understanding the specific needs and challenges faced by students can lead to more effective strategies for improving success rates.

Further research is needed to explore the complex interplay of variables affecting student success, including qualitative studies on student experiences and longitudinal analyses to track outcomes over time. However, it's essential to recognize the limitations of our method in capturing the full spectrum of factors influencing student success.

Appendix A: Method

Data was downloaded as a comma-separated (.csv) file and imported into Jupyter Notebook. The data was cleaned and prepared for analysis, focusing on the selection of relevant variables and addressing missing data. Multivariable Logistic regression was conducted using Python's statistical and machine learning libraries to identify significant predictors of student success in the preliminary year.

The initial step involved scrutinizing the dataset for missing values, especially in key predictor variables that could influence the analysis. Missing values within key predictor variables were addressed using imputation techniques for minor gaps and exclusion for variables with substantial missing information. Categorical variables were appropriately encoded to ensure their inclusion in the logistic regression model. The selection of predictor variables was guided by theoretical relevance and empirical evidence from preliminary data exploration.

Variables showing high multicollinearity were identified using the Variance Inflation Factor (VIF) analysis to prevent redundancy in the model. Those exceeding a certain threshold were considered for removal or modification to ensure the inclusion of independent predictors. Continuous variables were assessed for linearity with the logit of the outcome, with transformations applied as needed to meet logistic regression assumptions.

A logistic regression model was built using the cleaned and prepared dataset. The model aimed to predict the likelihood of a student completing the preliminary year based on selected predictor variables. The model's performance was evaluated using accuracy measures, a confusion matrix, and the Area Under the Curve - Receiver Operating Characteristic (AUC - ROC) score to assess its discriminative ability.

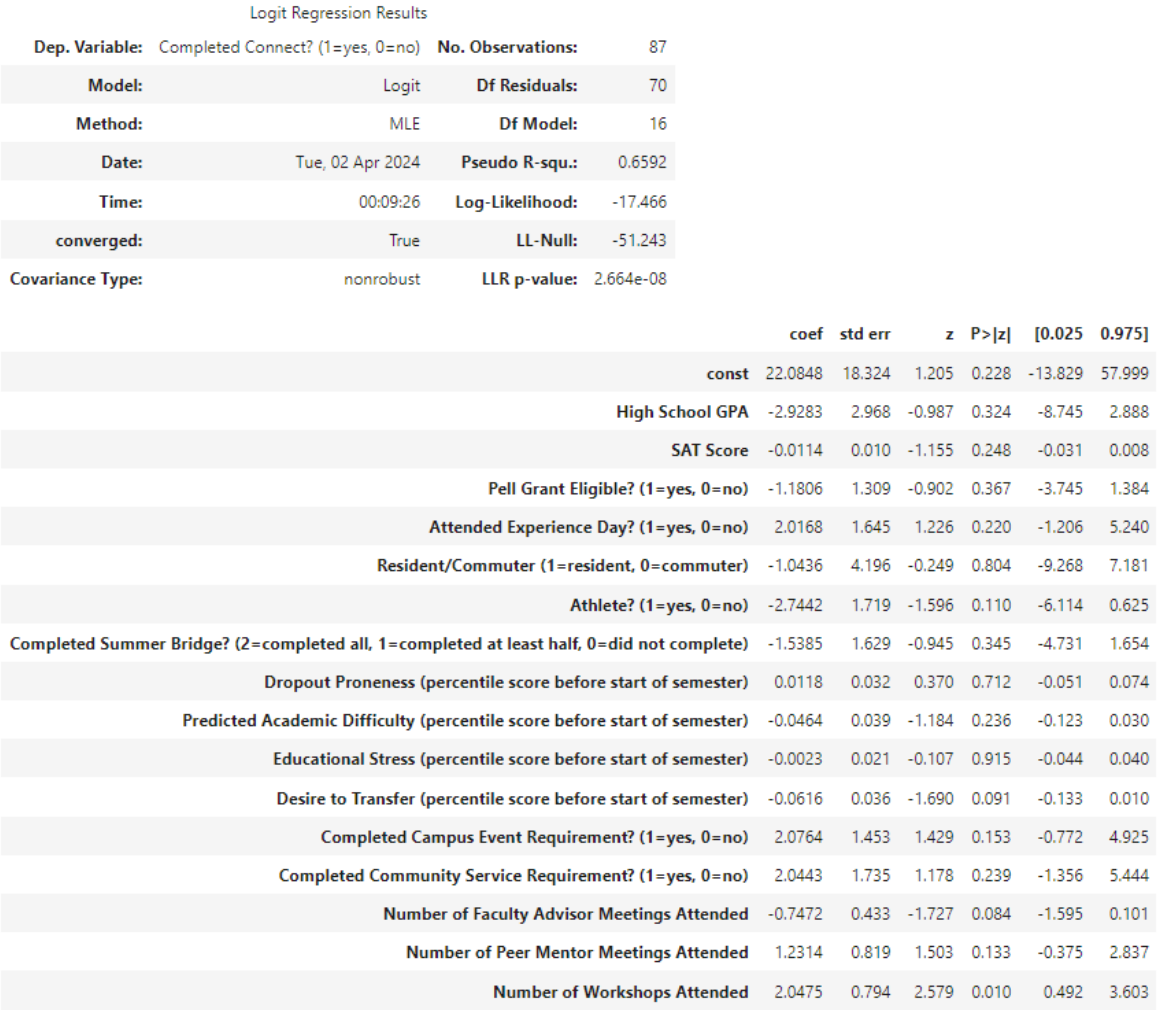
The methodology identified key predictors of student success, guiding targeted interventions to support students during the preliminary year and providing insights into the model's effectiveness.

Appendix B: Results

The CSV File contains 106 data points with 33 features to assess the student's success in the preliminary year. Out of this, 19 rows of missing values are removed from the dataset. Applied one hot encoding for categorical variables and removed irrelevant variables that are not directly related to student success in the preliminary year, such as Receptivity columns and variables with high multicollinearity to ensure robustness and interpretability.

Correlation Matrix was plotted to identify multicollinearity and variables showing high multicollinearity were identified using the Variance Inflation Factor (VIF) analysis to prevent redundancy in the model. Those variables that have a VIF value greater than 5 are removed to ensure smooth prediction.

The analysis employed the Logistic Regression algorithm from the statsmodel library providing quantitative insights into the factors influencing students' likelihood of completing the preliminary year. This model summary highlights the key findings emphasizing the role of various predictors.



Logit Regression Results

Dep. Variable:	Completed Connect? (1=yes, 0=no)	**No. Observations:**	87
Model:	Logit	**Df Residuals:**	70
Method:	MLE	**Df Model:**	16
Date:	Tue, 02 Apr 2024	**Pseudo R-squ.:**	0.6592
Time:	00:09:26	**Log-Likelihood:**	-17.466
converged:	True	**LL-Null:**	-51.243
Covariance Type:	nonrobust	**LLR p-value:**	2.664e-08

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	22.0848	18.324	1.205	0.228	-13.829	57.999
High School GPA	-2.9283	2.968	-0.987	0.324	-8.745	2.888
SAT Score	-0.0114	0.010	-1.155	0.248	-0.031	0.008
Pell Grant Eligible? (1=yes, 0=no)	-1.1806	1.309	-0.902	0.367	-3.745	1.384
Attended Experience Day? (1=yes, 0=no)	2.0168	1.645	1.226	0.220	-1.206	5.240
Resident/Commuter (1=resident, 0=commuter)	-1.0436	4.196	-0.249	0.804	-9.268	7.181
Athlete? (1=yes, 0=no)	-2.7442	1.719	-1.596	0.110	-6.114	0.625
Completed Summer Bridge? (2=completed all, 1=completed at least half, 0=did not complete)	-1.5385	1.629	-0.945	0.345	-4.731	1.654
Dropout Proneness (percentile score before start of semester)	0.0118	0.032	0.370	0.712	-0.051	0.074
Predicted Academic Difficulty (percentile score before start of semester)	-0.0464	0.039	-1.184	0.236	-0.123	0.030
Educational Stress (percentile score before start of semester)	-0.0023	0.021	-0.107	0.915	-0.044	0.040
Desire to Transfer (percentile score before start of semester)	-0.0616	0.036	-1.690	0.091	-0.133	0.010
Completed Campus Event Requirement? (1=yes, 0=no)	2.0764	1.453	1.429	0.153	-0.772	4.925
Completed Community Service Requirement? (1=yes, 0=no)	2.0443	1.735	1.178	0.239	-1.356	5.444
Number of Faculty Advisor Meetings Attended	-0.7472	0.433	-1.727	0.084	-1.595	0.101
Number of Peer Mentor Meetings Attended	1.2314	0.819	1.503	0.133	-0.375	2.837
Number of Workshops Attended	2.0475	0.794	2.579	0.010	0.492	3.603

Variables indicative of students' academic preparation, and engagement with campus support services, such as tutoring and counseling, emerged as significant predictors. Higher levels of engagement were positively associated with the successful completion of the preliminary year.

The model demonstrates a high accuracy rate of 92% in predicting student outcomes. Additionally, it achieves a significant Area Under the ROC Curve (AUC) score of 0.96, indicating its ability to distinguish between students likely to succeed and those who are not. For probabilities exceeding 0.5, students are predicted to pass, while those below 0.5 are predicted to fail. Out of 87 instances, only 7 student successes were incorrectly predicted. This suggests that the model effectively captures the relationship between predictor variables and the probability of success in the preliminary year.

The analysis offers insights for targeting interventions, like enhanced advising and mentoring programs. It underscores the need for understanding factors driving success, enabling UMass Dartmouth to tailor resources. This study of the logistic regression results provides a robust foundation for strategic decision-making, guiding interventions to enhance student success and retention.

Appendix C: Code

In this appendix, we document the Python code for performing logistic regression analysis on the dataset of students in the preliminary year program at UMass Dartmouth.

```
# Importing the libraries

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from statsmodels.stats.outliers_influence import variance_inflation_factor

import statsmodels.api as sm

import statsmodels.formula.api as smf

from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

from sklearn.metrics import precision_score, recall_score, f1_score, roc_auc_score
```

```
# Handling missing data, categorical data and unnecessary columns

data = pd.get_dummies(data, columns=['Gender','Federal Ethnic Group'])

data.dropna(inplace=True)

data = data.drop(['Reason for not Completing Connect', 'Reason not Retained', 'Retained F17-F18? (1=yes, 0=no)', 'Receptivity to Institutional Help (percentile score before start of semester)', 'Receptivity to Personal Counseling (percentile score before start of semester)', 'Receptivity to Academic Assistance (percentile score before start of semester)', 'Receptivity to Social Engagement (percentile score before start of semester)', 'Receptivity to Career Guidance ((percentile score before start of semester)', 'Receptivity to Financial Guidance (percentile score before start of semester)', 'F17 GPA', 'S18 GPA', 'CUM GPA', 'Number of Credits Earned', 'Attended Orientation? (1=yes, 0=no)', 'Gender_M','Gender_F','Federal Ethnic Group_Asian'], axis=1)
data['Completed Connect? (1=yes, 0=no)'] = data['Completed Connect? (1=yes, 0=no)'].replace('0, contract for fall', 1)

data['Completed Connect? (1=yes, 0=no)'] = data['Completed Connect? (1=yes, 0=no)'].astype(int)
```

```
# Handling Multicollinearity

predictors = data.drop(columns=['Completed Connect? (1=yes, 0=no)'])

# Calculate VIF for each predictor

vif_data = pd.DataFrame()

vif_data["Predictor"] = predictors.columns

vif_data["VIF"] = [variance_inflation_factor(predictors.values, i) for i in range(len(predictors.columns))]

# Remove predictors with VIF above a certain threshold (e.g., 5)

threshold = 5

collinear_vars = vif_data[vif_data["VIF"] > threshold]["Predictor"]

df = data

df.drop(columns=collinear_vars, inplace=True)
```

```
# Logistic Regression

x = df.drop('Completed Connect? (1=yes, 0=no)', axis=1)

y = df['Completed Connect? (1=yes, 0=no)']

model = sm.Logit(y, sm.add_constant(x))

results = model.fit( maxiter=5000)

results.summary()
```

```
# Evaluation metrics

pred = results.predict(sm.add_constant(x))

binary_pred = (pred >= 0.5).astype(int)

print('Accuracy: {:.2f}'.format(accuracy_score(y, binary_pred)))

print('f1_score: {:.2f}'.format(f1_score(y, binary_pred)))

print('AUC_ROC: {:.2f}'.format(roc_auc_score(y, pred)))

print("Confusion Matrix:\n", confusion_matrix(y, binary_pred))
```

Output:

Accuracy: 0.92
f1_score: 0.94
AUC_ROC: 0.96
Confusion Matrix:

```
 [[20  4]
 [ 3 60]]
```

Contribution:

Sindhuja Baikadi - 02128756: Worked on the Issues, Findings, Discussion, Method, and Results sections. Also self-plotted the graphs to analyze the data using the various methods discussed in the report.

Veda Sahaja Bandi - 02105111: Outlining key observations and insights, Analyzed the data thoroughly and worked on the coding portion of the project to implement necessary functionalities and features.