# Can We Predict College Admission? A Logistic Regression Analysis on Preliminary Year Students

## The issues:

The dataset used in the college success analysis was collected from a single institution in the United States, and it contains information on 107 students. The data was collected from several sources, including academic records, student surveys, and institutional records. The study was conducted by a team of researchers who were interested in understanding the factors that contribute to college students completing a preliminary year.

- What is the classification accuracy of the logistic regression in predicting student success?
- What are the important features in predicting student success in college?
- What are the implications of the findings for predicting and improving student success in college?
- What are the features that are affecting students from not success in college?
- What are the precision, recall, and F1-score of the selected algorithm in predicting student success?

## Findings:

- The logistic regression model has an accuracy of 0.955, meaning that it correctly predicted the class for 95.5% of the observations in the dataset.
- The most important feature in predicting success is the "Number of Credits Earned", with a coefficient of 0.245. This suggests that students who have earned more credits are more likely to be retained.
- The second and third most important features are "Number of Peer Mentor Meetings Attended" and "Number of Workshops Attended", with coefficients of 0.116 and 0.107, respectively. This implies that attending peer mentor meetings and workshops may positively impact retention.
- "Predicted Academic Difficulty" is the least important feature in predicting retention, with a coefficient of -0.028. This suggests that predicted academic difficulty does not have a significant impact on retention.
- The logistic regression model has high precision (1.0), which means that when it predicts a student will be retained, it is correct 100% of the time. However, its recall is 0.938, meaning that it correctly identifies 93.8% of the students who were retained. This indicates that the model is better at identifying true negatives than true positives.

## Discussion:

Based on the logistic regression analysis, several key findings have emerged regarding the factors that are most strongly associated with academic success at this institution.

- First, the number of credits earned is the strongest predictor of academic success, this suggests that students who can complete a higher number of credits are more likely to be successful academically.
- Second, the number of peer mentor meetings attended, and the number of workshops attended are also strong predictors of academic success. This indicates that students who take advantage of mentorship opportunities and attend workshops are more likely to succeed academically.
- Third, completing the Connect program is also positively associated with academic success. This suggests that the Connect program may be effective in supporting students and improving their academic outcomes.
- Fourth, several other factors, including F17 GPA, CUM GPA, and receptivity to social engagement, are also positively associated with academic success, although to a lesser degree.
- Finally, there are several factors that are negatively associated with academic success, including predicted academic difficulty, receptivity to institutional help, and desire to transfer, among others.

These findings have several implications for the institution. First, they suggest that providing students with support and mentorship opportunities, such as peer mentoring and workshops, can be effective in improving their academic outcomes. Second, the Connect program may be a valuable tool in supporting students and improving their academic success. Finally, efforts to address factors that are negatively associated with academic success, such as predicted academic difficulty and receptivity to institutional help, may be necessary to improve student outcomes.

# Appendix A: Method

## *Data collection.*

The dataset used in the college success analysis was collected from a single institution in the United States, and it contains information on 107 students. The data was collected from several sources, including academic records, student surveys, and institutional records. The study was conducted by a team of researchers who were interested in understanding the factors that contribute to college students completing a preliminary year.

## *Variable creation.*

here's a detailed description of the variables used in the college success analysis:

1. High School GPA: This is the student's high school grade point average at the time of admission to the college.
2. SAT Score: This is the student's score on the SAT exam, which is a standardized test used for college admissions in the United States.
3. Federal Ethnic Group: This is the ethnic group to which the student belongs as identified on their federal financial aid application.
4. Gender: This is the gender of the student.
5. Pell Grant Eligible? (1=yes, 0=no): This variable indicates whether or not the student is eligible for a Pell Grant, which is a need-based grant for low-income students in the United States.

6. Attended Orientation? (1=yes, 0=no): This variable indicates whether or not the student attended the college's orientation program for new students.
7. Attended Experience Day? (1=yes, 0=no): This variable indicates whether or not the student attended the college's experience day for prospective students.
8. Resident/Commuter (1=resident, 0=commuter): This variable indicates whether the student is a resident or commuter student.
9. Athlete? (1=yes, 0=no): This variable indicates whether the student is a student-athlete.
10. Completed Summer Bridge? (2=completed all, 1=completed at least half, 0=did not complete): This variable indicates whether the student completed the college's summer bridge program for incoming students.
11. Dropout Proneness (percentile score before start of semester): These variable measures the student's likelihood of dropping out of college based on their responses to a survey administered before the start of the semester.
12. Predicted Academic Difficulty (percentile score before start of semester): These variable measures the student's predicted academic difficulty based on their responses to a survey administered before the start of the semester.
13. Educational Stress (percentile score before start of semester): These variable measures the student's level of educational stress based on their responses to a survey administered before the start of the semester.
14. Receptivity to Institutional Help (percentile score before start of semester): These variable measures the student's receptivity to institutional help based on their responses to a survey administered before the start of the semester.
15. Receptivity to Academic Assistance (percentile score before start of semester): These variable measures the student's receptivity to academic assistance based on their responses to a survey administered before the start of the semester.
16. Receptivity to Personal Counseling (percentile score before start of semester): These variable measures the student's receptivity to personal counseling based on their responses to a survey administered before the start of the semester.
17. Receptivity to Social Engagement (percentile score before start of semester): These variable measures the student's receptivity to social engagement based on their responses to a survey administered before the start of the semester.
18. Receptivity to Career Guidance ((percentile score before start of semester): These variable measures the student's receptivity to career guidance based on their responses to a survey administered before the start of the semester.
19. Receptivity to Financial Guidance (percentile score before start of semester): These variable measures the student's receptivity to financial guidance based on their responses to a survey administered before the start of the semester.
20. Desire to Transfer (percentile score before start of semester): These variable measures the student's desire to transfer to another college based on their responses to a survey administered before the start of the semester.
21. Completed Campus Event Requirement? (1=yes, 0=no): This variable indicates whether the student completed the college's campus event requirement.
22. Completed Community Service Requirement? (1=yes, 0=no): A binary variable indicating whether the student completed the community service requirement.
23. Number of Faculty Advisor Meetings Attended: The number of meetings the student had with their faculty advisor.

24. Number of Peer Mentor Meetings Attended: The number of meetings the student had with their peer mentor.
25. Number of Workshops Attended: The number of workshops the student attended.
26. F17 GPA: The GPA of the student in the Fall 2017 semester.
27. S18 GPA: The GPA of the student in the Spring 2018 semester.
28. CUM GPA: The cumulative GPA of the student.
29. Number of Credits Earned: The total number of credits the student earned.
30. Completed Connect? (1=yes, 0=no): A binary variable indicating whether the student completed the Connect program.
31. Reason for not Completing Connect: A categorical variable indicating the reason why the student did not complete the Connect program.
32. Retained F17-F18? (1=yes, 0=no): A binary variable indicating whether the student was retained from Fall 2017 to Spring 2018.
33. Reason not Retained: A categorical variable indicating the reason why the student was not retained.

## *Analytic Methods*:

- In the above analysis, median imputation was used to handle missing values in the dataset. Specifically, for each feature with missing values, the median value of that feature was calculated from the non-missing values and used to replace the missing values. This approach was chosen as it is a simple and commonly used imputation method that helps preserve the overall distribution of the data.
- In this analysis, logistic regression was used to analyze the data. Logistic regression is a statistical method used to model the probability of a binary outcome, for given a set of predictor variables. It is commonly used in machine learning, statistics, and social sciences to model relationships between predictor variables and a binary outcome.
- The logistic regression model used in this analysis involves fitting a linear regression equation to the log odds of the outcome variable. The model then uses the logistic function to transform the linear regression equation into a probability value, which is bounded between 0 and 1. The logistic function is characterized by an S-shaped curve, and it is used to model the relationship between the predictor variables and the outcome variable.
- In this analysis, the logistic regression model was used to predict the likelihood of a student being retained or not retained, based on a set of predictor variables. The model was trained on a dataset of historical student data, and then used to predict retention outcomes for a new set of students.
- The performance of the logistic regression model was evaluated using metrics such as accuracy, precision, recall, and F1 score. These metrics provide a measure of how well the model performs at correctly predicting the outcome variable and are used to evaluate the effectiveness of the model.
- Additionally, feature importance was calculated based on the magnitude of the regression coefficients, to determine which variables were most strongly associated with student retention.

## Appendix B: Results:

Based on the multiple regression analysis, we found that several variables were significantly associated with first-year retention rates.

- The most important predictor was the number of credits earned in the first year of college, with a positive coefficient of 0.245. This suggests that students who earn more credits in their first year are more likely to be retained for their second year.

- Other significant predictors included the number of peer mentor meetings attended (0.116), the number of workshops attended (0.107), and completed Connect program (0.044). These results suggest that students who actively engage in programs designed to support them, such as peer mentorship and workshops, are more likely to be retained.
- Furthermore, completed community service requirement (0.017) and completed campus event requirement (0.013) were also found to be significant predictors of retention. These findings suggest that extracurricular activities may have a positive impact on retention.
- On the other hand, some variables were negatively associated with retention. These included predicted academic difficulty (-0.028), receptivity to institutional help (-0.023), desire to transfer (-0.015), receptivity to financial guidance (-0.019), and high school GPA (-0.013). These results suggest that students who struggle academically and are less receptive to institutional support may be at higher risk for not returning for their second year.
- It is important to note that some variables that were expected to have an impact on retention, such as SAT scores and Pell Grant eligibility, were not found to be significant predictors in this analysis.
- Overall, these findings suggest that colleges and universities should focus on providing programs and support services that promote academic success and engagement outside of the classroom. Additionally, identifying and supporting students who are at higher risk for not returning, such as those with predicted academic difficulty and lower high school GPAs, may also be important for improving retention rates.

# Appendix C: Data and code:

In [1]:
```
import pandas as pd

# load data into a pandas dataframe
df = pd.read_csv('/Users/maram/Library/Containers/com.microsoft.Excel/Data/Desktop/Preliminary college year.csv')
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
```

In [2]:
```
df.head()
```

Out[2]:

| | High School GPA | SAT Score | Federal Ethnic Group | Gender | Pell Grant Eligible | Attended Orientation | Attended Experience Day | Resident/Commuter | Athlete | Completed Summer Bridge | Dropout Proneness | Predicted Academic Difficulty | Educational Stress | Rece Institu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2.595 | 970.0 | White | F | 0.0 | 1.0 | 0.0 | 1.0 | 1.0 | 2.0 | 55.0 | 86.0 | 33.0 | |
| 1 | 2.637 | 1020.0 | White | F | 0.0 | 1.0 | 0.0 | 1.0 | 1.0 | 2.0 | NaN | NaN | NaN | |
| 2 | 2.803 | 930.0 | Black/African American | F | 1.0 | 1.0 | 1.0 | 1.0 | 0.0 | 1.0 | 81.0 | 89.0 | 95.0 | |
| 3 | 2.354 | 1080.0 | Hispanic/Latino | M | 0.0 | 1.0 | 1.0 | 1.0 | 0.0 | 1.0 | 88.0 | 98.0 | 49.0 | |

fill NaN values with median

In [4]:
```
df.fillna(df.median(numeric_only=True).round(1), inplace=True)
df.head()
```

Out[4]:

| | High School GPA | SAT Score | Pell Grant Eligible | Attended Orientation | Attended Experience Day | Resident/Commuter | Athlete | Completed Summer Bridge | Dropout Proneness | Predicted Academic Difficulty | Educational Stress | Receptivity to Institutional Help | Receptivity to Academic Assistance | Rece Pe Cour |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2.595 | 970.0 | 0.0 | 1.0 | 0.0 | 1.0 | 1.0 | 2.0 | 55.0 | 86.0 | 33.0 | 53.0 | 25.0 | |
| **1** | 2.637 | 1020.0 | 0.0 | 1.0 | 0.0 | 1.0 | 1.0 | 2.0 | 80.0 | 77.0 | 75.0 | 65.0 | 55.0 | |
| **2** | 2.803 | 930.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.0 | 1.0 | 81.0 | 89.0 | 95.0 | 71.0 | 72.0 | |
| **3** | 2.354 | 1080.0 | 0.0 | 1.0 | 1.0 | 1.0 | 0.0 | 1.0 | 88.0 | 98.0 | 49.0 | 52.0 | 76.0 | |
| **4** | 2.850 | 880.0 | 1.0 | 1.0 | 1.0 | 0.0 | 0.0 | 2.0 | 28.0 | 72.0 | 4.0 | 66.0 | 96.0 | |

drop the 'Reason for not Completing Connect' and 'Reason not Retained' are different for different person, so drop it for some time so that we can evaluate

In [5]:
```
df = df.drop('Reason for not Completing Connect', axis=1)
df = df.drop('Reason not Retained', axis=1)
```

In [6]:
```
df.head()
```

Out[6]:

| | Receptivity to Social Engagement | Receptivity to Career Guidance | Receptivity to Financial Guidance | Desire to Transfer | Completed Campus Event Requirement | Completed Community Service Requirement | Number of Faculty Advisor Meetings Attended | Number of Peer Mentor Meetings Attended | Number of Workshops Attended | F17 GPA | S18 GPA | CUM GPA | Number of Credits Earned | Completed Connect | Retained F17-F18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ) | 94.0 | 60.0 | 50.0 | 82.0 | 0.0 | 1.0 | 5.0 | 2.0 | 2.0 | 0.660 | 2.20 | 1.788 | 18.0 | 0.0 | 0.0 |
| ) | 48.0 | 55.0 | 61.0 | 63.0 | 0.0 | 0.0 | 3.0 | 1.0 | 2.0 | 1.567 | 1.66 | 1.625 | 18.0 | 0.0 | 0.0 |
| ) | 48.0 | 55.0 | 56.0 | 52.0 | 1.0 | 1.0 | 2.0 | 2.0 | 3.0 | 2.425 | 2.25 | 2.338 | 24.0 | 1.0 | 0.0 |
| ) | 25.0 | 65.0 | 61.0 | 96.0 | 0.0 | 0.0 | 2.0 | 2.0 | 0.0 | 0.925 | 2.60 | 2.600 | 6.0 | 0.0 | 0.0 |
| ) | 85.0 | 50.0 | 50.0 | 36.0 | 1.0 | 1.0 | 9.0 | 3.0 | 3.0 | 3.200 | 1.68 | 2.250 | 18.0 | 1.0 | 1.0 |

Split the data into training and testing sets to evaluate the performance of the logistic model.

In [7]:
```
from sklearn.model_selection import train_test_split

# split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(df.drop('Retained F17-F18', axis=1), df['Retained F17-F18'].astype(
```

Fit a logistic regression model to the training data, using a subset of the variables that are most likely to be significant predictors of the outcome based on domain knowledge or exploratory analysis.

In [8]:
```
df.head()
from sklearn.linear_model import LogisticRegression
X = df.iloc[:,0:28]

df['Retained F17-F18'] = df['Retained F17-F18'].astype(int)
y = df['Retained F17-F18']

# fit a logistic regression model
model = LogisticRegression(max_iter=8000,solver='sag')
model.fit(X,y)
```

Out[8]:
```
LogisticRegression(max_iter=8000, solver='sag')
```

Evaluate the performance of the model

In [9]:
```
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

print(f'Accuracy: {accuracy:.3f}')
print(f'Precision: {precision:.3f}')
print(f'Recall: {recall:.3f}')
print(f'F1 score: {f1:.3f}')
```

```
Accuracy: 0.955
Precision: 1.000
Recall: 0.938
F1 score: 0.968
```

In logistic regression, the coefficient measures the change in the log-odds of the response variable for a one-unit change in the predictor variable, holding all other variables constant. A larger magnitude of the coefficient indicates a stronger association between the predictor variable and the response variable. Therefore, a variable with a large positive or negative coefficient is more likely to be an important predictor.

In [10]:
```
# Identify important features
coefficients = pd.DataFrame({"feature": X.columns, "coefficient": model.coef_[0]})
coefficients = coefficients.sort_values(by="coefficient", ascending=False)
print("Important features:\n", coefficients)
```

```
Important features:
                                    feature  coefficient
26                    Number of Credits Earned     0.245291
21      Number of Peer Mentor Meetings Attended     0.115661
22                 Number of Workshops Attended     0.107480
27                           Completed Connect     0.044153
23                                    F17 GPA     0.043904
25                                    CUM GPA     0.017987
14            Receptivity to Social Engagement     0.017553
19      Completed Community Service Requirement     0.017385
8                           Dropout Proneness     0.016759
18          Completed Campus Event Requirement     0.012968
12          Receptivity to Academic Assistance     0.012299
7                       Completed Summer Bridge     0.011217
3                          Attended Orientation     0.010222
20  Number of Faculty Advisor Meetings Attended     0.007189
4                       Attended Experience Day     0.006803
13          Receptivity to Personal Counseling     0.005817
24                                    S18 GPA     0.004856
10                         Educational Stress     0.003478
15              Receptivity to Career Guidance     0.001132
6                                     Athlete     0.000713
2                         Pell Grant Eligible    -0.002026
1                                   SAT Score    -0.003571
5                             Resident/Commuter    -0.010196
0                               High School GPA    -0.012960
17                           Desire to Transfer    -0.014957
16           Receptivity to Financial Guidance    -0.018616
11           Receptivity to Institutional Help    -0.022991
9                Predicted Academic Difficulty    -0.028275
```

## References:

(i) Reference: “An Introduction to Statistical Learning with Applications in R”, Chapter 4, pages 129-196
(ii) Additional reference: Applied Logistic Regression, Hosmer & Lemeshow
(iii) sklearn https://scikit-learn.org/stable/
(iv) pandas https://pandas.pydata.org/