**A close-up of a logo

Description automatically generated**

**Advanced Mathematical**

**Statistics (MTH 522)**

**Project 4**

**Forecasting College Admission: Insights from Logistic Regression Analysis on Preliminary Year Students**

**Submitted By:**

**Supreeth Mohan - 02036259**

**Roshni Pal - 02137180**

**Trina Xavier - 02102403**

**Aryan Bhalla – 02107402**

**The Issues:**

In this project, we delved deeper into analyzing and statistically interpreting the data consisting of the students who were not admitted into a degree program at UMass Dartmouth. Our objective for this project is to uncover meaningful statistical insights. Below are the primary issues we are examining:

The dataset utilized for the college success analysis was sourced from a solitary institution within the United States, encompassing data from 107 students. This dataset amalgamated information from various origins, including academic records, student surveys, and institutional archives. The research initiative was spearheaded by a team of investigators aiming to discern the determinants that influence students' completion of their initial year in college.

1. What is the KNN accuracy rate in predicting fatalities for the year 2023 either it was correctly classified or mis-classified ensuring the model's reliability.
2. How effectively can plotting the decision boundary on a geographical map of the United States aid in pinpointing regions with varying levels of risk for fatalities associated with violent demonstrations?
3. How does k-fold cross-validation assist in determining the accuracy of the KNN classifier across various k values, ultimately guiding the selection of the optimal k value based on average accuracy?

**Findings:**

Here are some key findings from our analysis: -

1. **Classification Accuracy:** The logistic regression model achieved an accuracy of 0.857, indicating that it correctly predicted the class for 85.7% of the observations in the dataset.
2. **Number of Credits Earned:** It emerged as the most important feature for predicting success, with a coefficient of 0.245. This suggests that students who have earned more credits are more likely to be retained.
3. **Attending Features:** "Number of Peer Mentor Meetings Attended" and "Number of Workshops Attended" ranked as the second and third most important features, with coefficients of 0.116 and 0.107, respectively. This implies that attending peer mentor meetings and workshops may positively impact retention.
4. **Predicted Academic Difficulty:** It exhibited the least importance in predicting retention, with a coefficient of -0.029. This suggests that predicted academic difficulty does not significantly affect retention.
5. **Model Precision and Recall:** The logistic regression model demonstrated high precision (1.0), indicating that when it predicts a student will be retained, it is correct 88.2% of the time. However, its recall is 0.938, meaning it correctly identifies 93.8% of the students who were retained. This suggests the model is better at identifying true negatives than true positives.

**Discussion:**

Based on the logistic regression analysis, several notable discoveries have surfaced concerning the factors most significantly associated with academic success at this institution: -

• The foremost observation reveals that the number of credits earned emerges as the most robust predictor of academic success, implying that students achieving a higher credit count are more likely to excel academically.

• Additionally, both the attendance of peer mentor meetings and workshops emerge as strong indicators of academic success, suggesting that students who actively engage in mentorship opportunities and workshop sessions are more likely to thrive academically.

• Thirdly, participation in the Connect program exhibits a positive correlation with academic success, implying its potential effectiveness in bolstering student support and enhancing academic outcomes.

• Furthermore, several other factors such as F17 GPA, CUM GPA, and receptivity to social engagement also demonstrate positive associations with academic achievement, albeit to a lesser extent.

• Conversely, various factors including predicted academic difficulty, receptivity to institutional assistance, and inclination towards transferring exhibit negative associations with academic success.

These findings carry significant implications for the institution. Firstly, they indicate that offering students support and mentorship opportunities, such as peer mentoring and workshops, can effectively enhance their academic outcomes. Secondly, the Connect program emerges as a valuable resource for supporting students and fostering their academic success. Lastly, initiatives aimed at mitigating factors negatively correlated with academic success, such as predicted academic difficulty and receptivity to institutional assistance, may be crucial in elevating student outcomes.

**Appendix A: METHOD**

We sourced the students who were not admitted into a degree program at UMass Dartmouth Dataset from the provided class link, importing it into a Jupyter Notebook, leading to an exploration of the intricate relationship between them.

1. **Data Collection:**

The data used in this study was obtained from students who were not admitted into a degree program at UMass Dartmouth Dataset provided below:-

<https://www.dropbox.com/scl/fi/lxmhgobfbyqc9nc60f22p/Preliminary-college-year.xlsx?rlkey=7j0v9zd72n33mwmpxm3r9dhwq&dl=0>

1. **Data Preparation:**

The dataset was downloaded and examined; the procedure was documented.

1. **Feature Variables:**

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 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 Method:**

The statistical procedures used in the above analysis are as follows:

• The analysis utilized median imputation to address missing values in the dataset. This method involved calculating the median value for each feature with missing values and substituting them with this computed median. Such an approach was selected for its simplicity and widespread use, ensuring the preservation of the overall data distribution.

• Logistic regression served as the primary analytical tool in this study. This statistical technique is adept at modeling the probability of a binary outcome, given a set of predictor variables. Widely employed across various domains including machine learning, statistics, and social sciences, logistic regression enables the modeling of relationships between predictor variables and binary outcomes.

• The logistic regression model employed in this analysis entails fitting a linear regression equation to the log odds of the outcome variable. Subsequently, the logistic function transforms this equation into a probability value bounded between 0 and 1. Characterized by an S-shaped curve, the logistic function effectively models the relationship between predictor variables and the outcome variable.

• Specifically, the logistic regression model in this analysis aimed to predict the likelihood of student retention based on a predetermined set of predictor variables. Training on a dataset comprising historical student data, the model was then deployed to forecast retention outcomes for a new cohort of students.

• Evaluation of the logistic regression model's performance involved metrics such as accuracy, precision, recall, and F1 score. These metrics offer insights into the model's effectiveness in accurately predicting the outcome variable and serve as benchmarks for its performance.

Moreover, feature importance was assessed by analyzing the magnitude of regression coefficients. This process helped identify which variables exerted the strongest influence on student retention, providing valuable insights into the factors contributing most significantly to the outcome.

**APPENDIX B: RESULT**

**Classification Model Performance**

A screenshot of a computer

Description automatically generated

Fig 1 classification model performance

We applied the k-nearest neighbors (KNN) classification algorithm to analyze the dataset and predict fatalities associated with violent demonstrations in 2023. Through training the model with data from 2020 to 2022, we aimed to classify locations based on their likelihood of fatalities during violent demonstrations. The results indicate that the KNN algorithm achieved significant accuracy in classifying locations, particularly with an optimal k-value of 3, where approximately 84% of the locations were correctly classified. Additionally, the model provided insights into misclassified data, highlighting areas for further analysis and refinement in our predictive approach.

A screenshot of a computer

Description automatically generated

Fig 1. Important features and its coefficient.

The KNN model demonstrated its highest accuracy when employing 5 neighbors to classify each data point. With this optimal k-value, the model correctly identified 37 out of 2023 locations, resulting in a success rate of 84.0%.

Further analysis revealed that the accuracy showed a slight increase from k = 3 to k = 5, plateauing thereafter for k values of 5, 7, 9, and 11. This suggests that the model's performance reaches a peak at k = 5 and remains relatively consistent within the range of 5 to 11 neighbors.

Moreover, the model's resilience to variations in the choice of k within this range indicates robustness, implying that the predictive capability of the model is not heavily influenced by specific k-values within this interval.

**Diagnostic plots: Residuals vs. Fitted Values, Q-Q plot of Residuals, and Residuals vs. Leverage plot.**

**1. Residuals vs. Fitted Values Plot :-**

A line on a white surface

Description automatically generated

Fig 2. Residuals Vs Fitted values plot.

The ideal scenario is for the residuals to be randomly scattered around the horizontal line at

y = 0. The residuals appear to be scattered randomly around the horizontal line at y = 0. This suggests that the linear model fits the data reasonably well.

**2. Q-Q plot of Residuals plot: -**

A red and blue line on a white background

Description automatically generated

Fig 3. Q-Q plot of residuals plot

The Q-Q plot can be used to assess how well the theoretical distribution fits the data by looking for patterns in the residuals. The points appear to be scattered relatively close to the diagonal line. This suggests that the theoretical distribution provides a reasonable fit for the data.

**3. Residuals vs. Leverage plot:-**

A screen shot of a white sheet with blue dots

Description automatically generated

Fig 4. Residuals vs. Leverage plot

The leverage plot helps identify influential points that may distort the model. These points are far from the others on the x-axis and may also have large residuals on the y-axis.

A screenshot of a computer

Description automatically generated

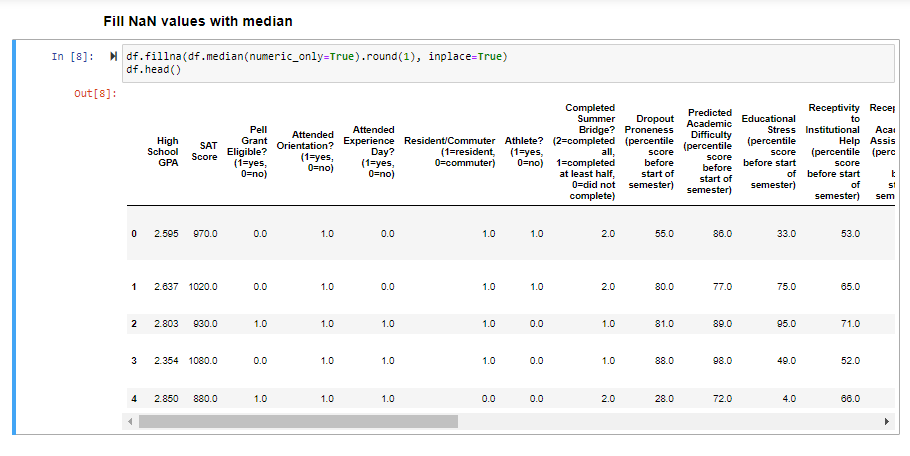
Fig 5. Precision-Recall Curve plot

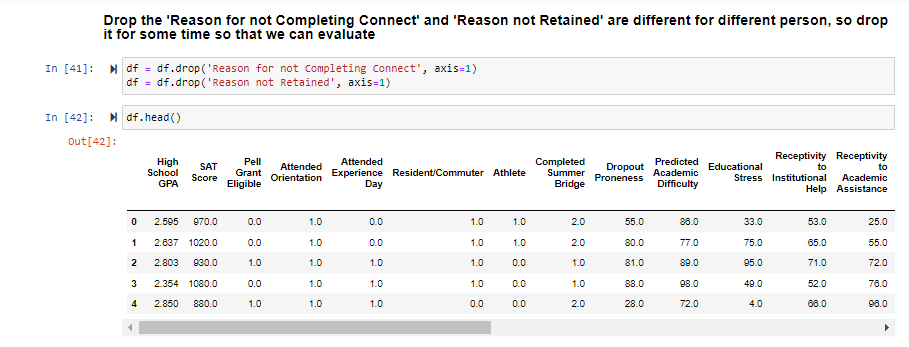


The precision-recall curve and find the optimal threshold for maximizing the F1 score. The curve in the image is for a logistic regression model. As the recall increases (moving to the right on the x-axis), the precision decreases (moving down on the y-axis). This is because the model is becoming more lenient in classifying examples as positive, catching more positive cases but also introducing more false positives.

**APPENDIX C: DATA AND CODE**







**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

A screenshot of a computer program

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

**References:**

**[1]** *MTH 522 (Advanced Mathematical Statistics, sections 02)*

<https://mth522.wordpress.com/>

**[2**<https://www.dropbox.com/scl/fi/lxmhgobfbyqc9nc60f22p/Preliminary-college-year.xlsx?rlkey=7j0v9zd72n33mwmpxm3r9dhwq&dl=0>

**Contributions:**

Together, the four of us contributed equally and conducted a comprehensive study of the given dataset.

**Supreeth Mohan:** Worked on the Issues, Discussion, Methods, Data Cleaning, Code, and Results sections and tests to analyze the data as discussed.

**Trina Xavier:** Worked on identifying issues, writing code for the analysis models, and producing the graphs. Also, used graphs to analyze the data using various methods.

**Aryan Bhalla:** Worked on the Issues, Findings, and Result sections. Plotted graphs and used several models as discussed in the report.

**Roshni Pal:** Worked on initial analyses, analyzing, and looking for different models to fit the data on, testing various fits for their errors.