PREDICTING STUDENT GRADUATION & DROPOUT

Authored by

Yonatan Navon *David Goldstein* *Ariel Friedlander*

Sep. 2024, Bar Ilan University

Link to GitHub repository

INTRODUCTION & ABSTRACT

At some point in their academic career, every student was burdened with the task of handing in a college or university application – likely in the multitude rather than one often. The simple reason – admission.

The thing about academic institutions is that for the most part, their admissions tend to be a process not intended to prevent prospective students from attaining an education, but rather uphold a certain level of academic integrity while maintaining the reputation of said institution.

This process usually involves a deliberation regarding a prospective students' achievements as a means of guarantee that a student would fit the institution's goals / standards.

One of the more important factors in the ultimate settlement of the deliberation, is asking the question of "will this student fold under the pressure? Will they succeed? Or will they drop out?".

In that regard, we present a study which examines a set of factors that influence student success, centered around the question of "Will a student drop out?".

We analyze a dataset containing a selection of variables such as demographic, academic, and socio-economic – and perform said analysis with an assortment of statistical tools such as logistic regression, t-tests, non-parametric tests, and interaction terms.

Key findings through our research indicate that academic performance (grades) particularly though the first semester is a notably strong predictor of our main question of dropping out, while interaction effects between age as well as parental qualifications and scholarship status also play a role in prediction.

To back our findings, we trained a logistic regression model which achieves an AUC score of 0.85, demonstrating good overall classification performance. Furthermore, the implications of Type I and Type II errors are also discussed in the analysis.

METHODS

The analysis is based on a dataset containing 4,432 student records, each with 35 variables related to academic, demographic, and socio-economic factors.
The dependent binary variable "Target" indicates whether a student graduated (1) or dropped out (0).

The following methods were employed:

1. **Logistic Regression**: A binary classification model was built to predict student outcomes. Interaction terms between age and first-semester grades were included to explore combined effects. The model's performance was evaluated with an AUC score of 0.85.

- **Hypotheses**:
  - H_0: The coefficients for predictors are equal to zero (no effect on graduation probability).
  - H_1: The coefficients for predictors are non-zero (significant effect on graduation probability).

2. **T-tests**: Used to compare the means of continuous variables such as grades and age between graduates and dropouts.

- **Hypotheses**:
  - H_0: No significant difference in means between graduates and dropouts for the variable in question.
  - H_1: Significant difference in means between graduates and dropouts.

3. **Mann-Whitney U Test**: Applied to compare non-parametric distributions, such as parental qualifications.

- **Hypotheses:**
  - H_0: No significant difference in the distribution of parental qualifications between graduates and dropouts.
  - H_1: Significant difference in the distribution of parental qualifications.

4. **Kruskal-Wallis Test**: Used to evaluate differences in grades across different application modes.

- **Hypotheses:**
  - H_0: No significant difference in grades across different application modes.
  - H_1: Significant difference in grades across different application modes.

5. **Chi-Square Test**: Employed to assess the association between categorical variables such as scholarship status and graduation outcomes.

- **Hypotheses:**
  - H_0: No association between scholarship status and graduation outcomes.
  - H_1: Significant association between scholarship status and graduation outcomes.

6. **Bonferroni Correction**: Applied to account for multiple comparisons across hypothesis tests to reduce the risk of Type I errors.

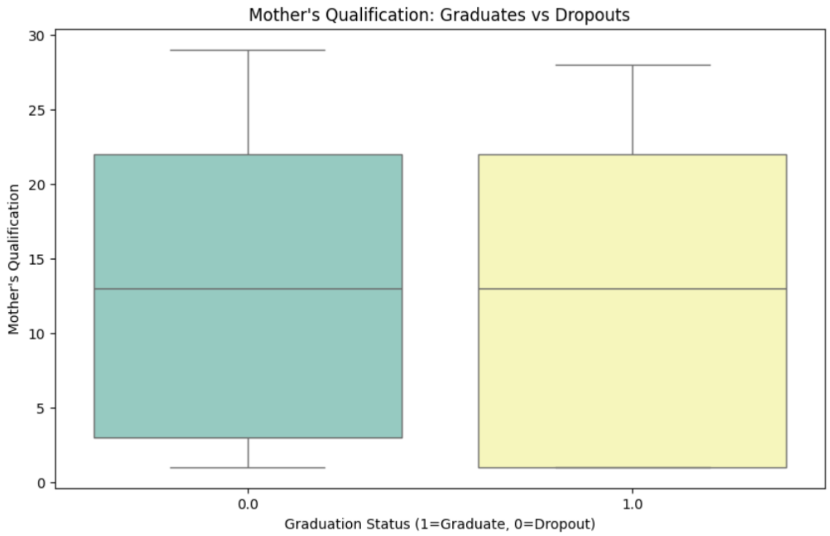
7. **Power Analysis**: Conducted for T-tests to assess the likelihood of Type II errors, with power values of 1.0 indicating no risk of failing to detect significant effects.

8. **ROC Curve and AUC**: The model's classification performance was evaluated using the ROC curve and an AUC of 0.85, demonstrating strong predictive power.

RESULTS

1. **Mann-Whitney U Test** for Mother's Qualification:

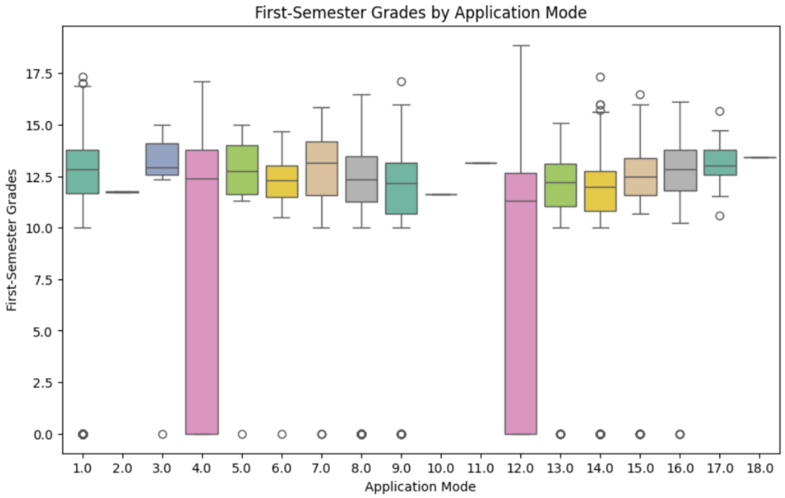
- Statistic: 556,804.0, P-value: 0.0848
- Result: No significant difference in ***mother's qualifications*** between graduates and dropouts.



It is easy to see that the boxes (showing the range of 2.5%-97.5% student's mother qualification) are very similar, therefore, we can see that the mother's qualification does not have much effect on the graduate/dropout rate.

2. **Kruskal-Wallis Test** for First-Semester Grades by Application Mode:

- Statistic: 720.1534, P-value: 3.869e-151
- Result: Significant differences in ***first-semester grades*** across ***application mode***



The boxplot shows the distribution of ***first-semester grades*** for each ***application mode*** (x-axis) with ***grades*** on the y-axis.

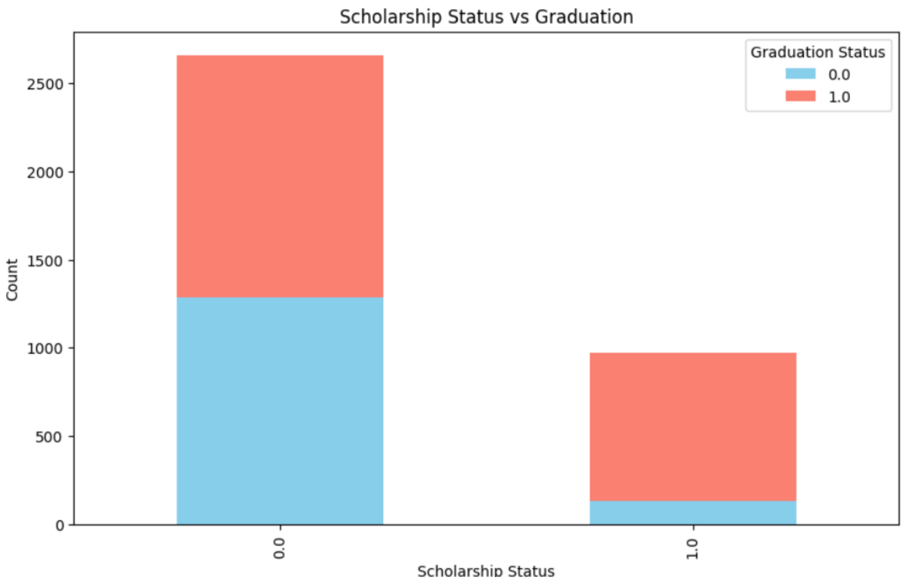
Each box represents the middle 50% of students' grades for a mode, with the line inside showing the **median**.

The whiskers extend to the range of most grades, and dots represent ***outliers***.

Modes 4.0 and 12.0 have very low and consistent grades, while others, like ***1.0, 2.0, and 6.0***, show wider distributions and higher median scores, indicating more variation in student performance. Outliers suggest some students performed exceptionally well or poorly in certain modes.

3. **Chi-Square Test** for Scholarship Status vs Graduation:

- Chi-Square Statistic: 364.783, P-value: 5.1091e-79
- Result: Strong association between scholarship status and graduation outcomes.

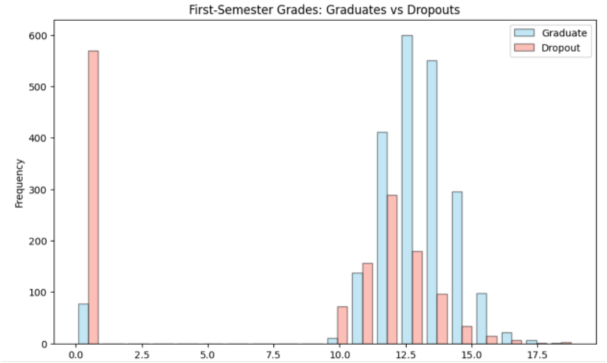


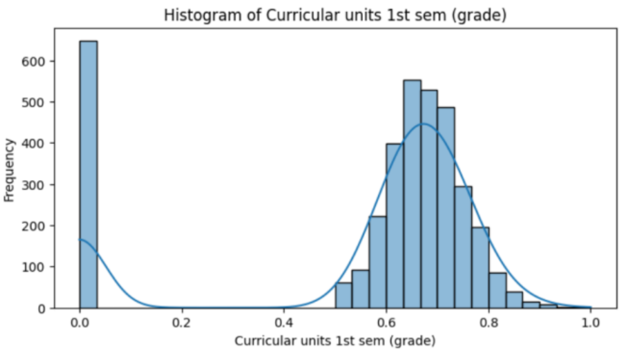
It can see that if the student got a Scholarship it effects strongly on the graduate/dropout rate, and who that didn't get a Scholarship is very likely to dropout

Please note that a student who received a scholarship can be affected by the university's forecast

4. **T-Test for First-Semester Grades** (Graduates vs Dropouts):

- Statistic: -32.755, P-value: ~0
- Result: Graduates have significantly higher first-semester grades compared to dropouts.

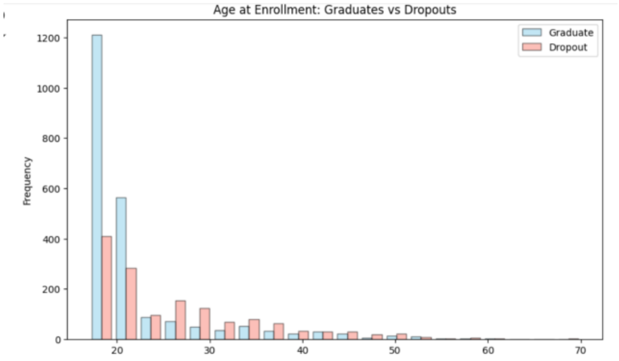


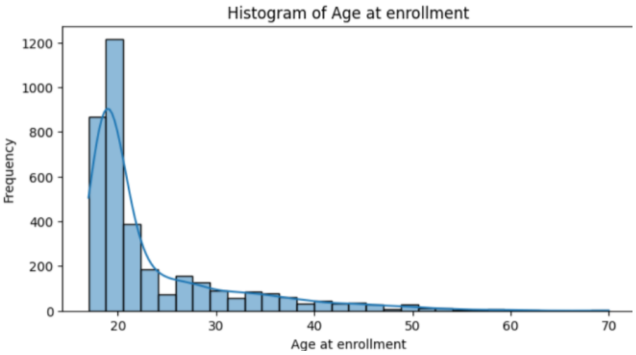


We see that students that fail first semester are very likely to dropout, compared to students who passed have about a third dropout rate, **with little dependence of the grade itself.**
Note: the value 0 is for all grades that are fails

5. **T-Test for Age at Enrollment** (Graduates vs Dropouts):

- Statistic: -15.548, P-value: ~0
- Result: Graduates tend to be younger than dropouts at the time of enrollment.

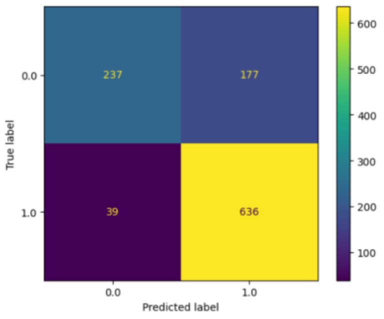


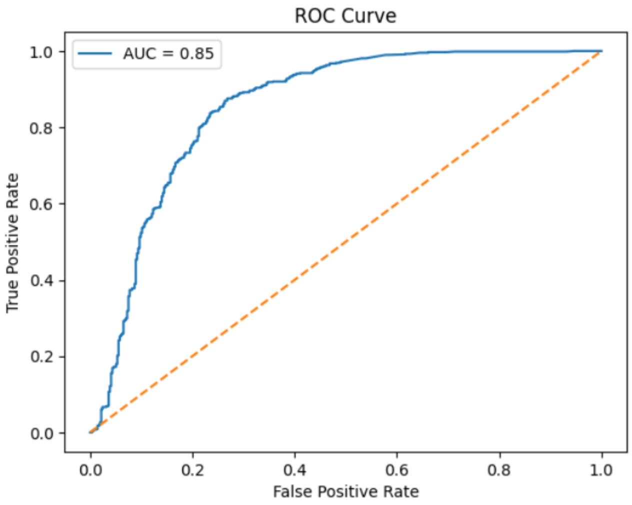


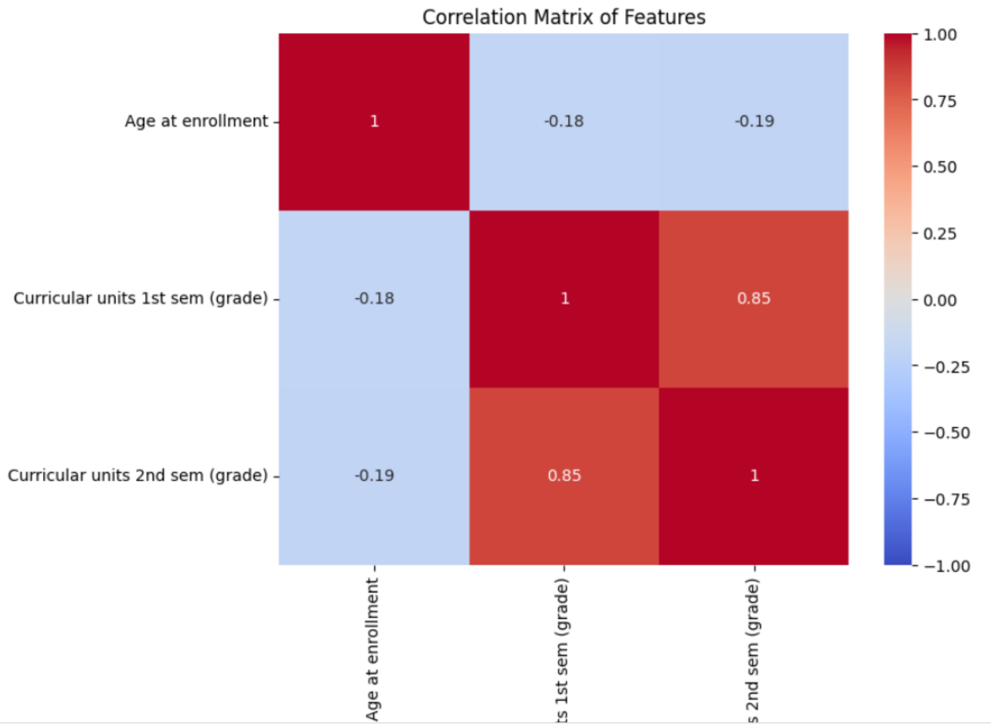
Easy to see it is best to be young 😊

6. **Logistic Regression** with Interaction Terms:

- **Age at Enrollment**: Coefficient: -0.3671, P-value: 0.0000
- **First-Semester Grades**: Coefficient: -0.5448, P-value: 0.0001
- **Interaction Term (Age x Grades)**: Coefficient: 0.0249, P-value: 0.0173
- AUC Score: 0.85 (indicating strong classification performance).







As you see we have a pretty good prediction, more on that in the conclusion

Discussion

The findings from the study affirm that academic performance, particularly first semester grades, plays a critical role in predicting whether a student will graduate or drop out.

- **First-Semester Grades**: Graduates had significantly higher first-semester grades than dropouts, as indicated by both the T-test and the logistic regression model.

- **Age at Enrollment**: Younger students were more likely to graduate, which was confirmed by the T-test results and the logistic regression coefficient.

- **Scholarship Status**: The Chi-Square test showed a strong association between being a scholarship holder and graduating, suggesting that financial aid plays a role in student success.

- **Interaction Effects**: The inclusion of interaction terms between age and first-semester grades revealed that while older students are generally less likely to graduate, this effect is less pronounced for those with higher first-semester grades.

- **Model Performance**: The logistic regression model achieved an AUC of 0.85, indicating good classification performance. However, the model's recall for dropouts was lower, suggesting that additional predictors or techniques (such as dropout-specific models) may improve sensitivity to at-risk students.

Conclusion

This study demonstrates that academic performance, particularly during the first semester, is the strongest predictor of whether a student will graduate or drop out.

Scholarship status and age at enrollment also play significant roles in predicting student success. The logistic regression model performed well, achieving an AUC of 0.85.

However, future improvements could focus on better identifying students at risk of dropping out through the inclusion of additional socio-economic variables or advanced machine learning techniques to increase recall for dropout prediction.