Alexis Hancz
QMSS 301 002
Sophie Geoghan
24 February 2023

Predicting Students' Academic Success: Modeling Multivariate Logistic Regression

Methods: This report investigates the extent to which an undergraduate student's second semester grade average, scholarship status, and gender predict their likelihood to graduate or drop out. UC Irvine's dataset (3630 observations of undergraduate students) forms the foundation of my multivariate logistic regression in Python. The dataset consists of 13 binary, numeric and categorical variables, ranging from parents' education levels to course of study. From these, I selected a student's second semester grade (x = *secsemgrade* (0-20)), scholarship status (x= *ownschshp*, no scholarship = 0, scholarship = 1), and gender (x = *gender*, male = 1, female = 0) to predict academic success (binary predictor, y = *target*, 0=dropped out, 1=graduated). I chose these variables out of interest and to maximize my model's accuracy.

I. First, I imported a variety of Python libraries using *pandas* to prepare my analysis.
II. Second, I imported UC Irvine's academic success dataset into Python as a CSV file ("dropout_data.csv") using *pandas* and explored the dataset through summary statistics and the response counts by variable (see **Appendix D**).
III. Third, I checked for imbalances in the dataset that could potentially skew my model using *numpy*.
IV. Fourth, I fitted a multivariate logistic regression model over my variables of interest using a train-test split strategy using *statsmodels* and *sklearn*. The model predicts the chances of dropout or graduation based on my three explanatory variables. To maximize my model's performance, I trained and tested it on portioned sample data (train size = 0.8, random state = 19). I trained it on 80% of the dataset (2904 data rows) to provide the model with enough data variability, and excluded 20% (726 data rows) to be tested on later. I created a new column "prob" in the dataset to store the trained model's predicted probabilities of whether a student will drop out or graduate; a probability close to one indicates a higher likelihood of graduation. I then plotted the predicted probabilities using *matplotlib* and *seaborn* (**Appendix A).** Finally, I created a new column "pred2true" that stores the predicted probabilities as binary outcomes (<0.5 ="drop out", >0.5= "graduate") to calculate model performance later.
V. Fifth, I calculated the accuracy, sensitivity, and specificity of the trained model using *sklearn* (see Discussion**)**, and visualized the predicted outcomes in a confusion matrix **(Appendix B).**
VI. Lastly, I tested the trained model and evaluated its performance on the test. I expected and received a slightly higher overall model performance rate for the trained data it was built on compared to the

previously unseen test data. I evaluated and visualized the model's test predictions through accuracy, sensitivity, specificity using *sklearn* (see Discussion), and a confusion matrix **(Appendix B)**.

Results and Analysis: The trained model returned coefficients for each explanatory variable, which indicate the magnitude of the variable's effect on the predictor variable (graduating or dropping out). As typical for logistic regression, I interpreted the coefficients through their odds ratios as follows:

- *secsemgrade:* A 1 unit increase in a student's second semester grade average indicates they are 1.34 times more likely to be a graduate than a dropout, holding all other variables constant.
- *ownschshp*: If a student has a scholarship, they are 4.72 times more likely to graduate than a student without a scholarship, holding all other variables constant.
- *gender*: The odds of a male student graduating are 0.518 times the odds of a female student graduating, meaning males are less likely to graduate than females, holding all other variables constant.

Discussion: In terms of overall performance, the trained model had a slightly higher accuracy of 81.40% (meaning the trained model correctly predicted a dropout or graduate outcome 81.40% of the time) compared to the test set accuracy of 80.03%. The trained model also had a slightly higher sensitivity of 95.51% (meaning the trained model correctly predicted a graduate outcome 95.51% of the time) compared to the test set sensitivity of 93.75%, and a slightly higher specificity of 59.67% (meaning the trained model correctly predicted a dropout outcome 59.67% of the time) compared to the test set specificity of 57.91%. It is expected that a model performs better on the data it was trained on than previously unseen test data. Overall, the results indicate the model is better at identifying which students will graduate than which will drop out.

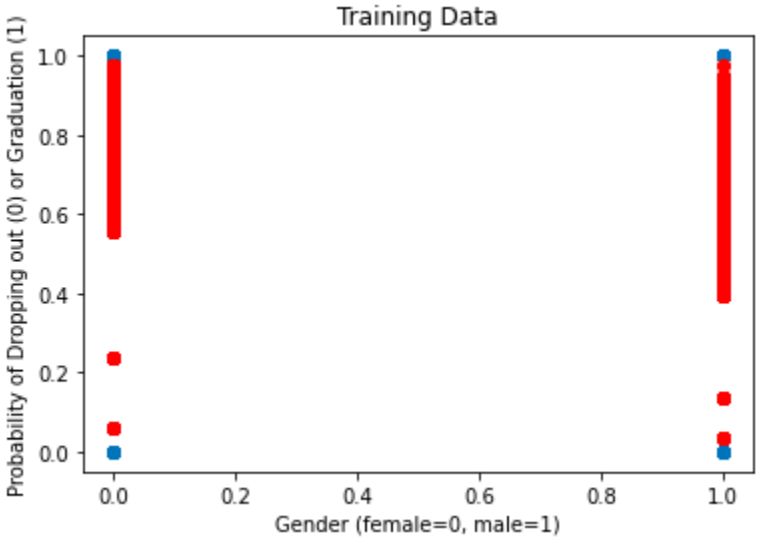
I used multiple strategies to maximize accuracy in my model. For one, I experimented with different variable combinations while building and training my model; for example, my initial trained model with gender, scholarship status, and educational special needs status yielded an accuracy of 68%, so I decided to replace educational special needs with students' second semester grades, and greatly boosted the model's accuracy. Secondly, I raised my random state value to correct for underfitting (rare, but did occur) which ultimately placed the trained model as slightly more accurate than the test.

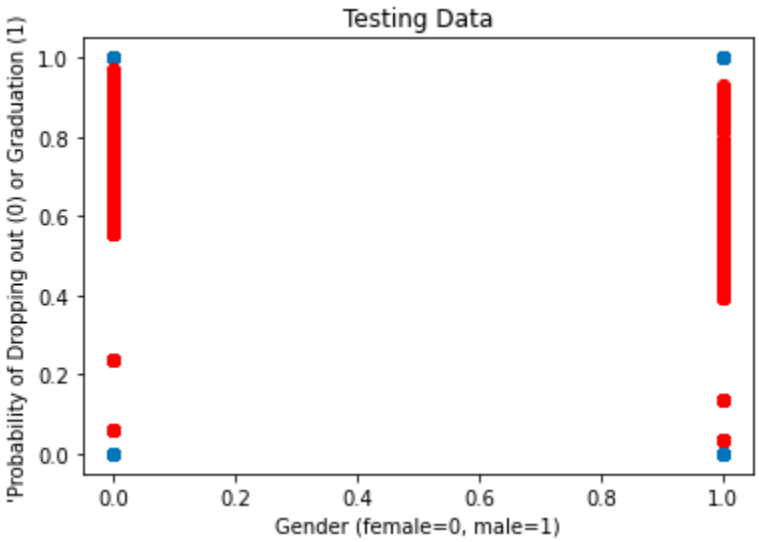
Conclusion: By prioritizing intentional variable selection and running a test-train strategy in a multivariate logistic regression, I was able to observe and analyze the effect of three diverse factors on the odds of whether a student will drop or graduate their undergraduate education. In terms of applications, other researchers would do well to investigate the underlying, qualitative connections between second semester grades, scholarship status, and gender that make them reasonable predictors of academic success when taken together. Ultimately, knowledge of the effect of these variables on academic success should be used to improve resources for students that bear dropout indicators during their time in undergrad.

Appendices

Appendix A

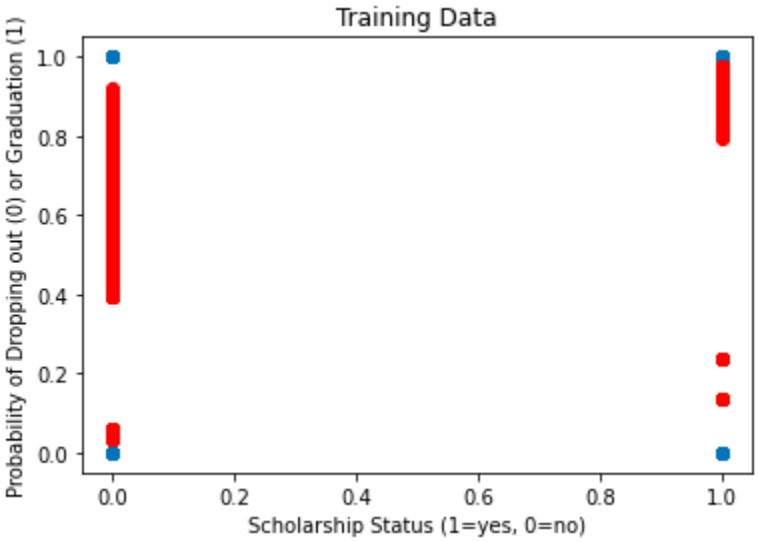
(i) Scatterplots of the predicted probabilities of graduation or dropout depending on whether a student is male or female. The blue dots represent the actual probability values, while the red dots represent the predicted probability values based on the training data. The condensed range of predicted probabilities near 1 for females indicates that being female predicts greater chances of graduation.





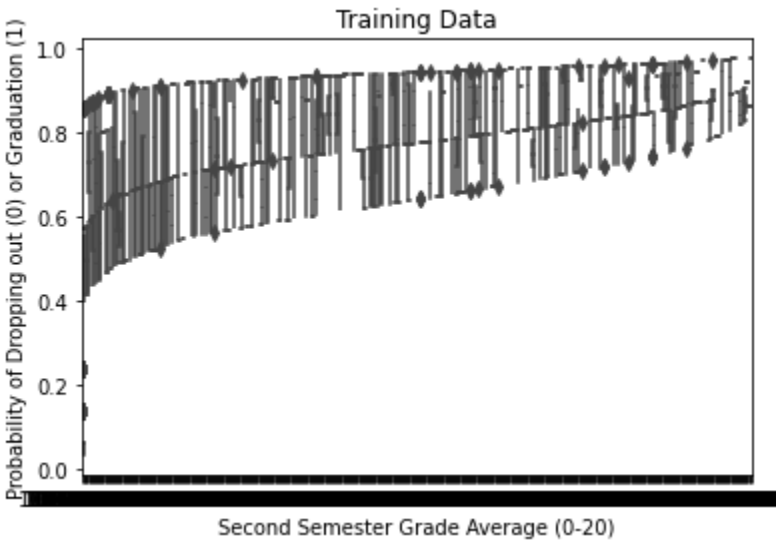
Appendix A

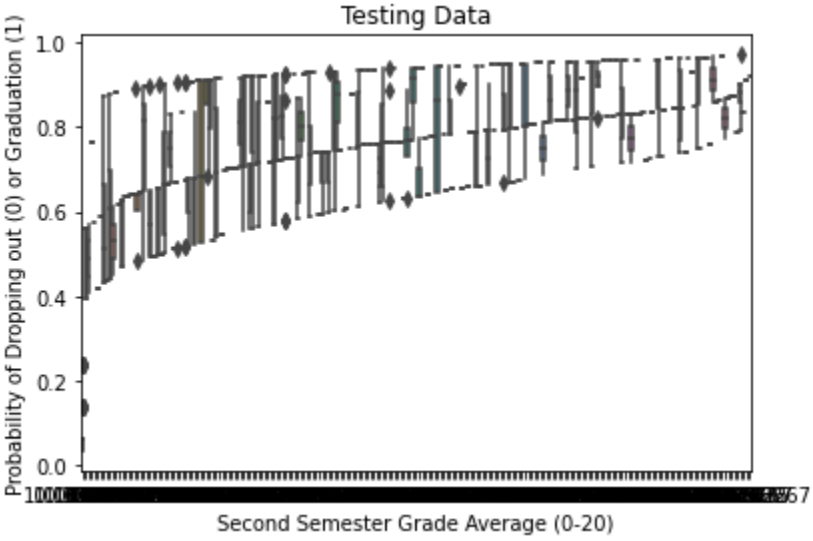
(ii) Scatterplots of the predicted probabilities of graduation or dropout depending on whether a student has a scholarship. The blue dots represent the actual probability values, while the red dots represent the predicted probability values based on the training data. The condensed range of predicted probabilities near 1 for having a scholarship indicates that having a scholarship predicts greater chances of graduation.



Appendix A

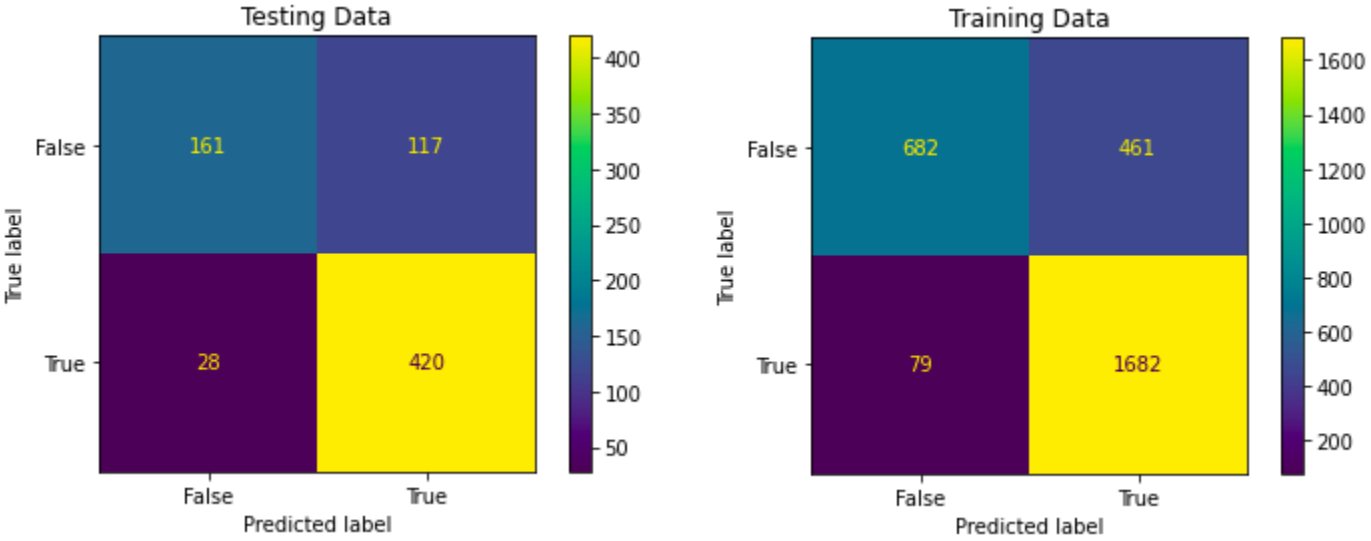
(iii) Plots of the predicted probabilities of graduation or dropout depending on a student's second semester grade average. The condensed range of predicted probabilities near 1 for higher grade averages indicates that having a higher second semester grade average predicts greater chances of graduation.





Appendix B

Confusion matrices summarizing model performance accuracies.



Appendix D: Basic information about the dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3630 entries, 0 to 3629
Data columns (total 15 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Unnamed: 0     3630 non-null   int64
 1   moqual         3630 non-null   int64
 2   faqual         3630 non-null   int64
 3   admingrade     3630 non-null   float64
 4   eduspecneeds   3630 non-null   int64
 5   prevgrade      3630 non-null   float64
 6   ownschshp      3630 non-null   int64
 7   prevqual       3630 non-null   int64
 8   displaced      3630 non-null   int64
 9   debtor         3630 non-null   int64
 10  paidfeetodate  3630 non-null   int64
 11  gender         3630 non-null   int64
 12  secsemgrade    3630 non-null   float64
 13  target         3630 non-null   int64
 14  course         3630 non-null   int64
dtypes: float64(3), int64(12)
memory usage: 425.5 KB
```