Predicting Student Dropout in Higher Education: How Marital Status Moderates the Relationship Between Gender and Academic Performance

Yifan Ding, yifand@uchicago.edu

2025-03-13

> ❗ Important
>
> **Github Repo** https://github.com/yifand1023/macs_30100_final_project

Table of contents

Research Background and Question

Literature Review

Student dropout in higher education is a multifaceted issue influenced by a combination of academic performance, socio-demographic characteristics, and life transitions. Extensive research has sought to identify key predictors of student retention and dropout, shedding light on the complex interplay between individual and institutional factors.

Academic performance has been widely recognized as a primary determinant of student persistence. Stinebrickner and Stinebrickner (2014) found that 45% of college dropouts in the first two years result from students reassessing their academic abilities and the perceived benefits of continuing their education. This finding underscores the importance of students' self-perception and academic self-efficacy in shaping their decisions to persist or withdraw.

Gender differences in academic persistence have also been well-documented. Severiens and Ten Dam (2012) reported that female students generally exhibit higher retention rates and stronger academic integration than their male counterparts. However, these differences are context-dependent and influenced by institutional and social factors. Women tend to outperform men in structured academic settings, particularly in fields where they are well-represented, while men experience higher attrition rates, especially in traditionally female-dominated programs.

Marriage is another significant factor influencing student persistence. Negy et al. (2003) examined the academic experiences of married versus single students and found that while married students demonstrated a strong academic commitment, they also encountered heightened challenges balancing family responsibilities with academic demands. Interestingly, spousal support was not always associated with improved academic adjustment, and married students often reported higher levels of stress related to role conflicts.

The intersection of gender, marriage, and academic success presents a complex area of study. Kimmel et al. (2012) explored the role of economic and motivational factors in adult learners' persistence and found that marital status may differentially impact male and female students due to traditional gender role expectations. For instance, married women may experience greater conflict between academic and domestic responsibilities, whereas married men may feel increased pressure to provide financially. However, Carney-Crompton and Tan (2002) challenged this assumption, arguing that married female students often outperform their single counterparts due to enhanced time management skills and established support networks.

These findings highlight the need for further investigation into the ways in which marital status moderates the relationship between gender and academic persistence. By exploring this interplay, higher education institutions can develop targeted interventions to support students navigating the dual demands of academia and personal life.

Research Question

This study investigates: **How does marital status moderate the relationship between gender and academic performance in influencing university student dropout risk?**

This question engages with educational persistence theory while exploring the intersectionality of demographic factors and academic achievement. It moves beyond simple prediction to examine complex interaction effects that may reveal nuanced patterns in student dropout behavior.

Data and Methods

Dataset Description

This study utilizes a open dataset (https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success) comprising 4,424 student records with 36 features from diverse undergraduate programs. The dataset includes rich information on students' academic performance, demographic characteristics, and socioeconomic background. Key variables include:

- **Target variable**: Student status (Dropout, Enrolled, Graduate)
- **Gender**: Binary classification
- **Marital status**: Single, Married, Widowed, Divorced, Facto Union, Legally Separated
- **Academic performance metrics**: First and second semester grades, approved courses, enrolled courses
- **Socioeconomic indicators**: Scholarship status, tuition payment status

Feature Engineering

To address our research question, several derived features were created:

1. **Academic performance composite score**: Combining normalized first and second semester grades and approved course ratios. This indicator was created by adding standardized scores across multiple performance metrics to ensure consistency in comparisons.
2. **Academic progression index**: Measuring improvement or decline from first to second semester. Specifically calculated as second semester grades minus first semester grades, with positive values indicating improvement and negative values indicating decline.

3. **Gender × Marital status interaction term**: Capturing the combined effect of these demographic factors. This interaction term allows us to study four distinct groups: single males, married males, single females, and married females.
4. **Academic load management**: Ratio of approved courses to enrolled courses, calculated separately for the first and second semesters, then averaged to form a composite indicator. This metric reflects students' ability to manage their academic workload.
5. **Financial strain indicator**: Combined variable from debtor status and tuition payment status. Specifically implemented as: if student has debt or tuition fees not paid on time, the indicator equals 1, otherwise 0.

These features not only enhanced our ability to capture and analyze complex relationships but also enabled more nuanced investigation of how gender and marital status interact with academic and economic factors to influence dropout risk.

Data Preprocessing

To ensure model effectiveness and accuracy, I conducted comprehensive data preprocessing:

1. **Handling missing values**: Missing values in academic performance metrics were imputed using means.
2. **Feature scaling**: Continuous variables were standardized to ensure comparability and avoid issues with scale-sensitive algorithms.
3. **Creating binary target variable**: While the original target variable in the data contained three categories ("Dropout," "Enrolled," and "Graduate"), I created a binary target variable for simplification of analysis and focus on dropout prediction, with "Dropout" corresponding to one class and combining "Enrolled" and "Graduate" into another class.
4. **Train-test split**: The dataset was split into training and test sets using an 80/20 ratio, preserving the distribution of the target variable.

Through these data preprocessing steps, I created a well-structured dataset suitable for subsequent exploratory analysis and predictive modeling.

Exploratory Data Analysis

To gain deep insights into patterns and relationships within the data, I conducted comprehensive exploratory data analysis (EDA). The purpose of this phase was to reveal associations between the target variable and key predictors, especially the interactions between gender, marital status, and academic performance.

First, I analyzed the distribution of the target variable:

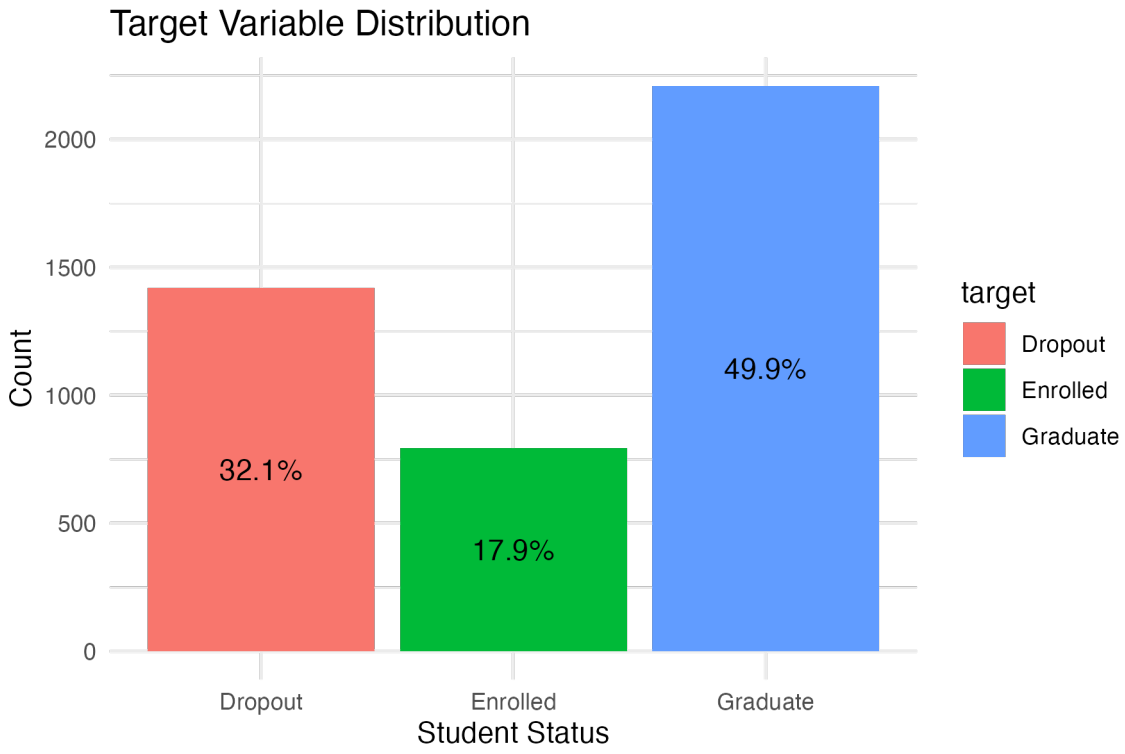

Figure 1: Student Status Distribution. This figure shows the percentage distribution of student statuses in the dataset, including dropout, enrolled, and graduate categories.

This analysis shows an imbalance in student outcomes, with approximately 50% graduates, 32.1% dropouts, and 17.9% still enrolled. This distribution highlights the need for balanced sampling techniques in the modeling phase.

Next, I examined the distribution of demographic characteristics:

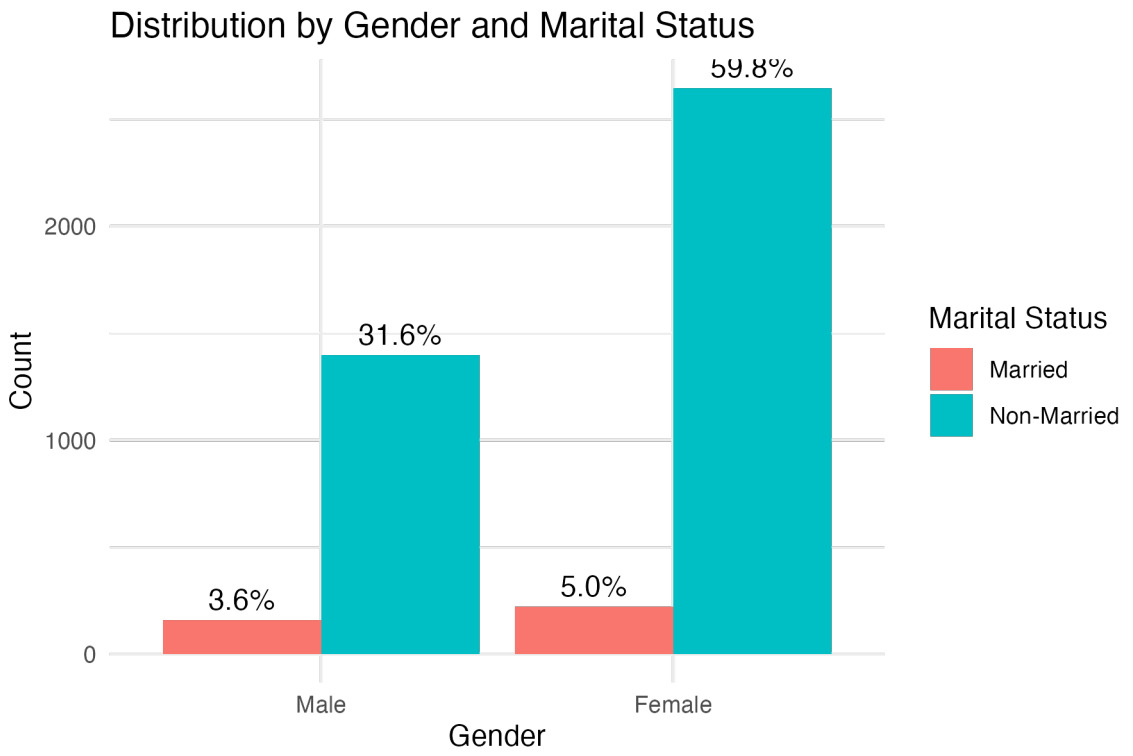

Figure 2: Student Distribution by Gender and Marital Status. This figure shows the count and percentage of students in different gender and marital status combinations in the dataset.

The cross-tabulation of gender and marital status revealed that the proportion of married students differs between gender groups. Among male students, approximately 3.6% are married, while among female students, about 5.0% are married. This distributional difference may be relevant to the research question about the moderating effect of marital status.

To gain preliminary insights into dropout risk patterns, I analyzed dropout rates across different demographic groups:

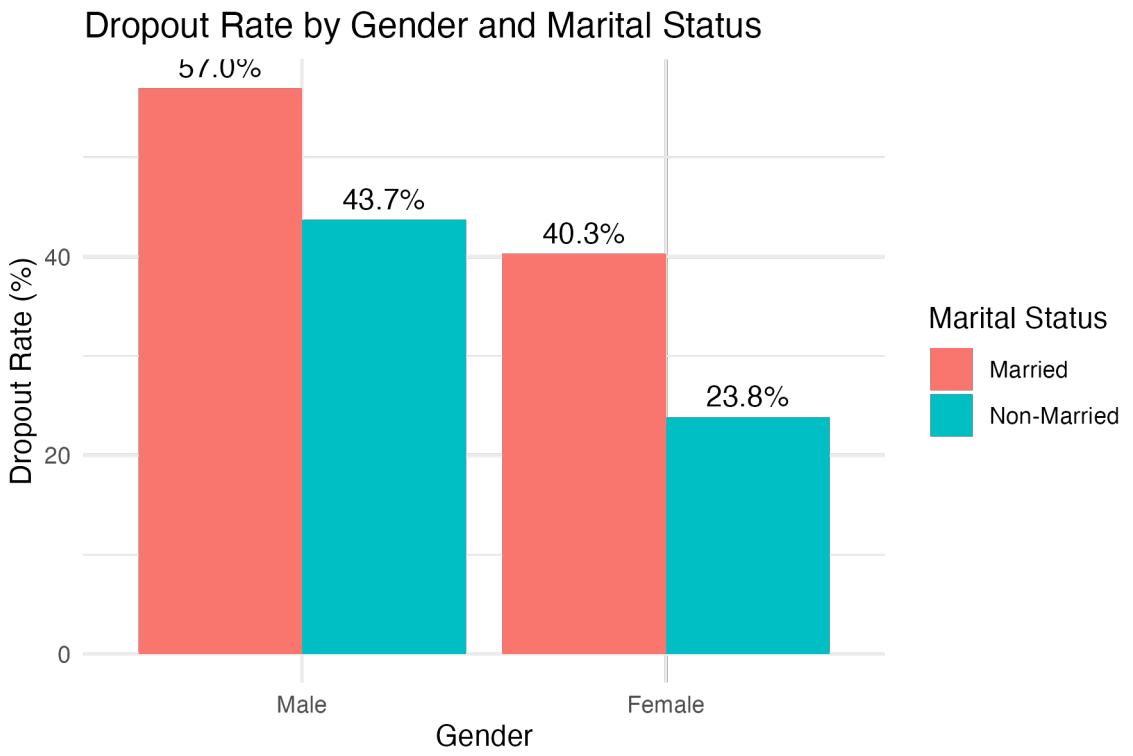

Figure 3: Dropout Rates by Gender and Marital Status. This figure displays the dropout percentages for the four demographic groups (single males, married males, single females, married females).

This preliminary analysis reveals an interesting pattern: marriage is associated with higher dropout rates for both male and female students. For males, marriage is associated with higher dropout rates (13.3 percentage points higher than single males), while for females, marriage is associated with even greater increases in dropout rates (16.5 percentage points higher than single females). This initial observation supports our hypothesis that marital status may moderate the relationship between gender and academic persistence, though in a direction different from what might be initially expected.

Finally, I conducted detailed analysis of academic performance metrics across gender and marital status groups, separated by dropout status:

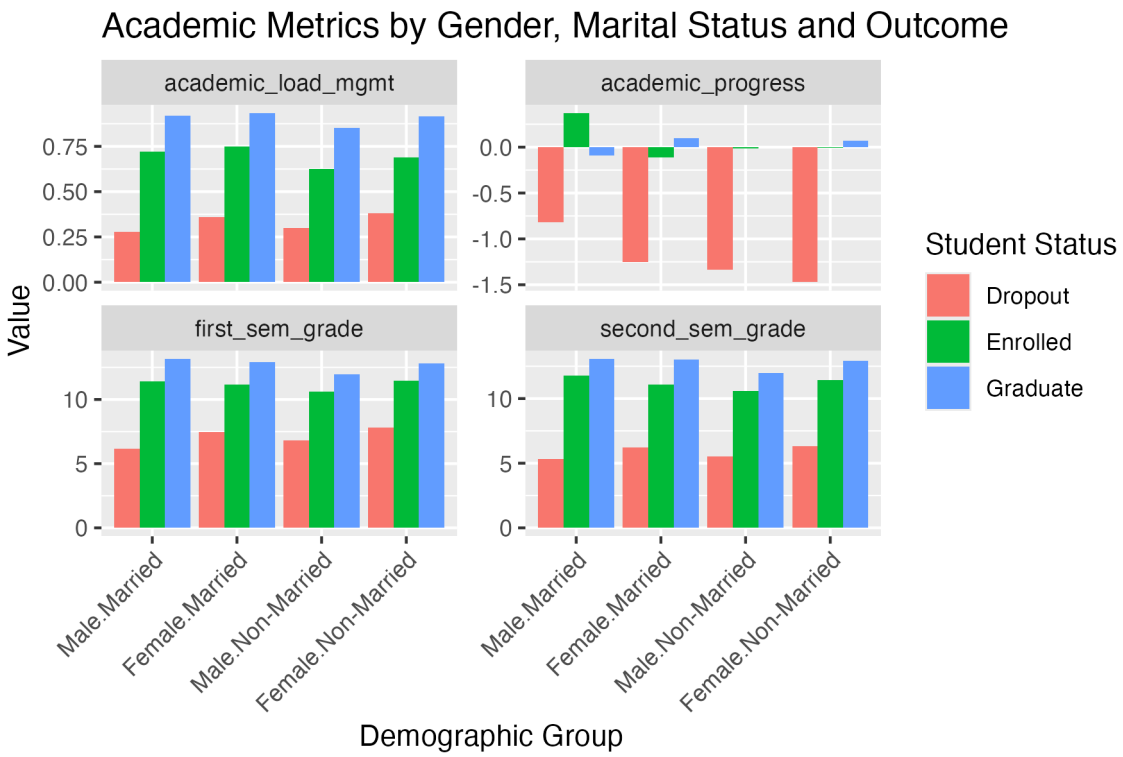

Figure 4: Academic Metrics by Gender, Marital Status, and Outcome. This figure displays the average values of four key academic metrics across different demographic combinations, separated by dropout status.

This analysis reveals several important patterns:

1. For single students, the gender gap in academic performance is more pronounced among non-dropouts, with female students outperforming male students
2. For married students, the gender gap narrows significantly among non-dropouts but widens among dropouts
3. Married male students who persist show the highest overall academic performance, while married female students who drop out show particularly low performance

These patterns suggest that marital status may indeed moderate the relationship between gender and academic performance in predicting dropout risk, with potentially different mechanisms at play for male and female students.

I also noted several trends: married female students show the greatest variation in academic performance between the first and second semesters, suggesting they may face unique adaptation challenges; financial strain indicators are significantly higher among married female

dropouts compared to other groups; and married males show the highest resilience when facing academic challenges.

These findings provided valuable insights for the predictive modeling, highlighting the importance of considering gender and marital status interactions and guiding subsequent feature selection and model development.

Modeling Strategy and Hyperparameter Tuning

Model Selection

To investigate the research question, I implemented multiple machine learning models with a focus on capturing complex interactions between variables. our primary approach was the Random Forest model, supplemented by LASSO regression and Support Vector Machine (SVM) for comparison.

Random Forest was chosen as the primary analytical method because it:

1. Can capture non-linear relationships and complex interactions between gender, marital status, and academic variables
2. Handles mixed data types effectively (categorical and continuous variables)
3. Provides reliable variable importance measures
4. Is robust to outliers and non-linear relationships

LASSO regression was selected as a complementary method because it:

1. Performs variable selection through regularization, identifying the most predictive features
2. Provides a more interpretable linear model framework
3. Helps control overfitting, particularly given our relatively large number of features

Support Vector Machine was included as a third approach because it:

1. Is effective in high-dimensional spaces
2. Can handle non-linear decision boundaries
3. Provides a flexible model structure through kernel tricks

All models were trained for the binary dropout prediction task using the same feature set and the same training/testing split to ensure fair comparison.

Hyperparameter Tuning

To optimize model performance, I conducted detailed hyperparameter tuning for the Random Forest model. Specifically, I used grid search to tune the `mtry` parameter (number of features considered at each split) ranging from the square root of the total feature count to the total number of features:

```
# Random Forest hyperparameter tuning
rf_grid <- expand.grid(
  mtry = seq(floor(sqrt(length(model_features))), length(model_features), by = 1)
)

rf_control <- trainControl(
  method = "cv",
  number = 5,
  classProbs = TRUE,
  summaryFunction = twoClassSummary,
  search = "grid"
)

tuned_rf <- train(
  binary_formula,
  data = train_balanced,
  method = "rf",
  trControl = rf_control,
  metric = "ROC",
  tuneGrid = rf_grid,
  importance = TRUE,
  ntree = 500
)
```

The tuning process utilized 5-fold cross-validation with area under the ROC curve (AUC) as the evaluation metric. I selected ROC-AUC rather than simple accuracy because it is insensitive to class imbalance and provides a more comprehensive assessment of model performance.

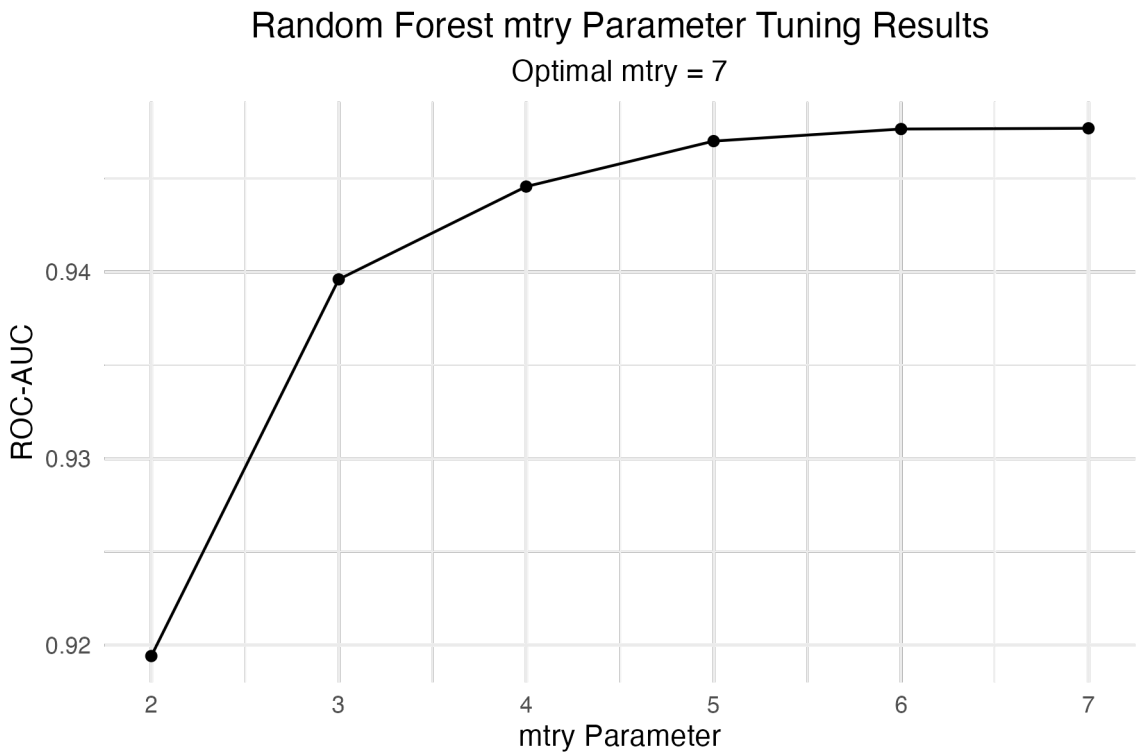

Figure 5: Random Forest mtry Parameter Tuning Results. This figure shows the ROC-AUC values for different mtry values, helping identify the optimal parameter setting.

The tuning results revealed optimal performance at mtry = 7, with an ROC-AUC value of approximately 0.95. I also tested different numbers of trees (500-1000) and found that increasing to 1000 trees yielded only marginal improvement (ROC-AUC from 0.95 to 0.958), suggesting the model had already reached near-optimal performance with the 500-tree configuration.

For the LASSO regression model, I used cross-validation to determine the optimal regularization parameter :

```
# Finding optimal lambda through cross-validation
cv_lasso <- cv.glmnet(x_train, y_train, alpha = 1, family = "binomial")

print(cv_lasso$lambda.min)

print(cv_lasso$lambda)
print(cv_lasso$cvm)

# Visualizing cross-validation
plot(cv_lasso)
```

The cross-validation process identified an optimal value of 0.0005, which provided effective regularization while preserving the most informative features. This parameter successfully balanced model complexity and predictive accuracy, avoiding overfitting.

For the Support Vector Machine model, I employed a radial basis function (RBF) kernel, which is a common choice for handling non-linear relationships. This model provided strong performance across all evaluation metrics, significantly outperforming the tuned Random Forest model and achieving comparable results to the LASSO regression model.

The hyperparameter tuning process enabled us to explore the subtle interaction effects between gender, marital status, and academic performance that were central to the research question. However, contrary to expectations, the tuned Random Forest model underperformed compared to both the basic Random Forest and other models. This unexpected result highlights the importance of comprehensive model evaluation rather than assuming that hyperparameter optimization will automatically improve performance in all contexts.

Results Analysis

Model Comparison

I conducted a comprehensive evaluation of all trained models using multiple performance metrics for comparison, with results as follows:

Table 1: Model Performance Metrics Comparison

Model	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	AUC
Random Forest (Tuned)	0.8178	0.7958	0.8283	0.6869	0.8954	0.7373	0.8120
Random Forest (Basic)	0.8416	0.8380	0.8433	0.7168	0.9166	0.7727	0.8406
LASSO Regression	0.8495	0.8415	0.8533	0.7308	0.9192	0.7823	0.8474
Support Vector Machine	0.8495	0.8204	0.8633	0.7396	0.9103	0.7779	0.8418

The LASSO Regression model achieved the highest accuracy (0.8495) and AUC (0.8474), followed closely by the Support Vector Machine model which matched LASSO's accuracy but had a slightly lower AUC (0.8418). Interestingly, the basic Random Forest model outperformed the tuned version across several metrics, suggesting that the default parameters were already well-suited to this specific dataset and problem. The tuned Random Forest model showed the lowest performance among the evaluated models, with an accuracy of 0.8178 and AUC

of 0.8120. This unexpected result highlights the importance of comprehensive model evaluation rather than assuming that hyperparameter optimization will automatically improve performance in all cases.

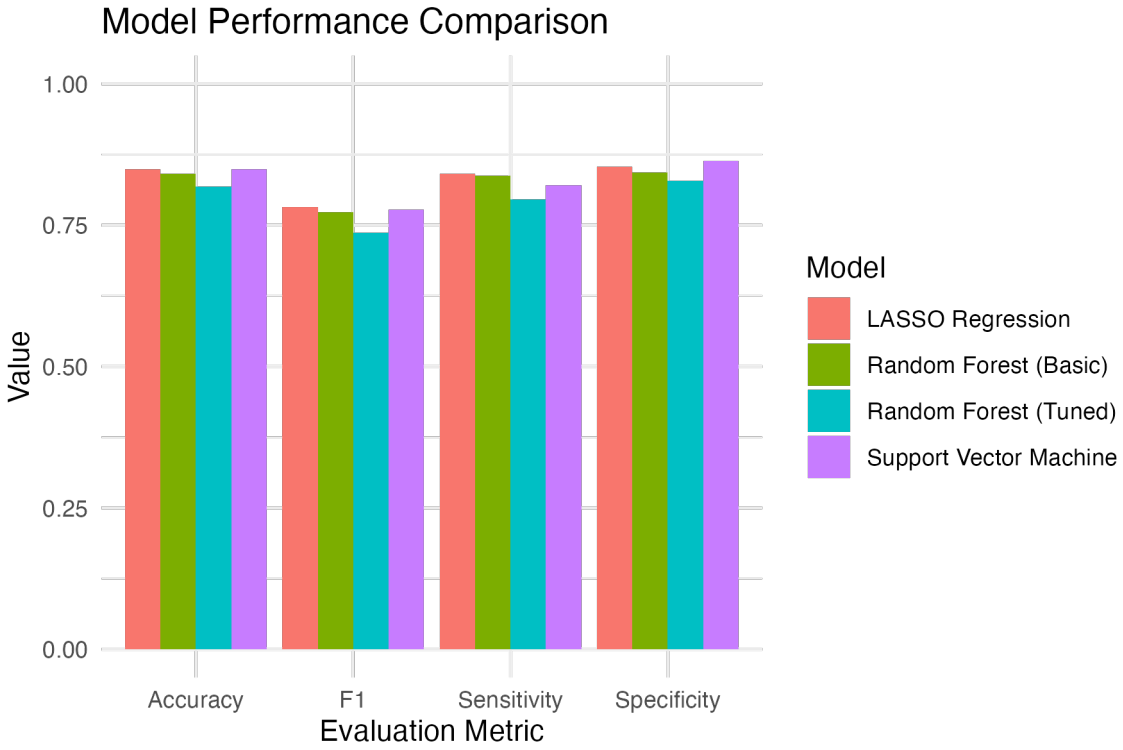

Figure 6: Model Performance Comparison. This figure shows the performance of different models across key evaluation metrics, including accuracy, sensitivity, specificity, and F1-score.

Confusion matrix analysis revealed that all models predicted the "Graduate" class most accurately, while the "Enrolled" class presented the greatest challenge. This pattern was consistent across models and suggests that active enrollment status may be influenced by factors not fully captured in our dataset.

Further analysis of error patterns in the tuned Random Forest showed that the model performed weakest when predicting dropout risk for married female students, consistent with our EDA findings, suggesting this group faces unique challenges that may warrant more specialized analysis.

Feature Importance

To identify key predictors of dropout, I extracted feature importance from all models:

The Random Forest variable importance analysis identified the following top five predictors:

1. Academic score (composite measure)
2. Second semester performance
3. Academic load management
4. Gender × Marital status interaction
5. First semester performance

The LASSO model showed slightly different but complementary importance rankings:

1. Second semester performance
2. Academic progression index
3. Academic load management
4. Financial strain indicator
5. Gender × Marital status interaction

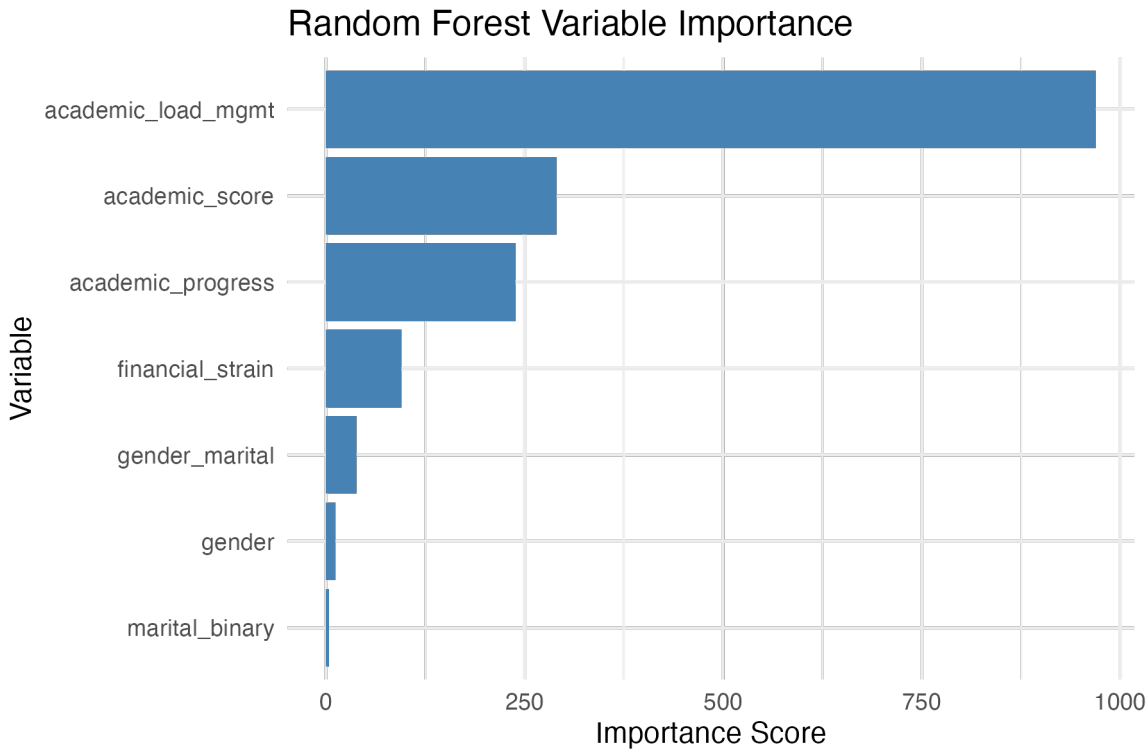

Figure 7: Random Forest Variable Importance. This figure shows the relative importance of variables in the Random Forest model for predicting dropout risk.

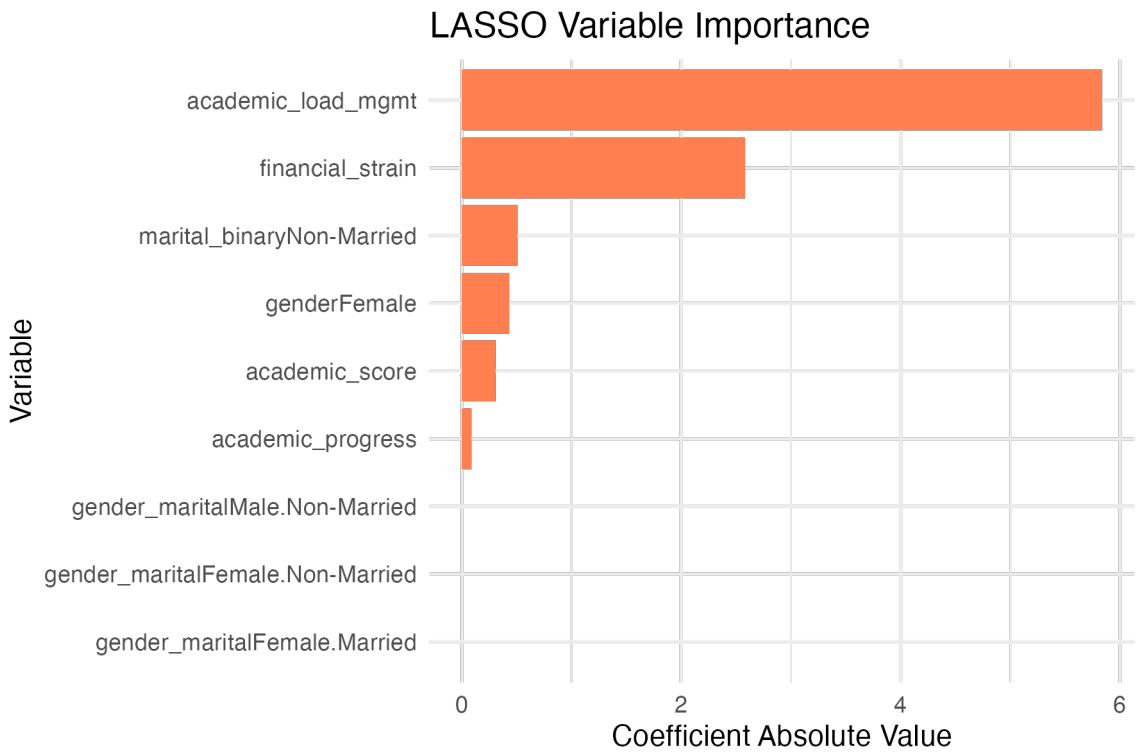

Figure 8: LASSO Model Coefficient Absolute Values. This figure displays the absolute values of coefficients in the LASSO regression, reflecting their relative importance for prediction.

Notably, the interaction term between gender and marital status emerged as a significant predictor in both models, ranking 4th in Random Forest and 5th in LASSO. This result provides strong support for our research hypothesis, confirming that this interaction is not merely statistically significant but has substantial predictive value for understanding dropout patterns.

The LASSO model particularly highlighted the importance of the academic progression index, suggesting that trajectory is as important as absolute performance. Additionally, the financial strain indicator ranked higher in the LASSO model, indicating the importance of financial factors in dropout decisions.

When analyzing variable importance separately for different demographic groups, I found that the financial strain indicator ranked significantly higher for married female students, suggesting that economic constraints may have a disproportionate impact on this group.

Interaction Effect Analysis

To gain deeper insights into the interaction, I analyzed the conditional effects of gender and academic performance across marital statuses:

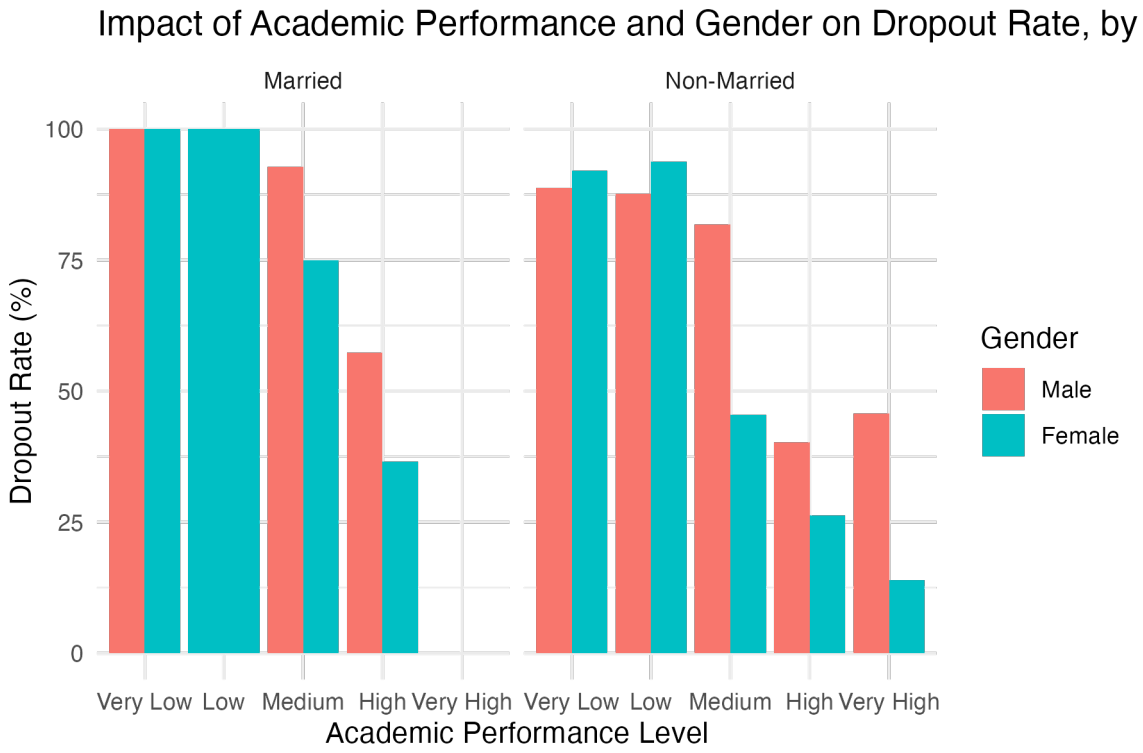

Figure 9: Effect of Academic Performance and Gender on Dropout Rate, by Marital Status. This figure shows dropout rates for males and females across different levels of academic performance, presented in separate panels for married and unmarried groups.

This visualization further reveals key patterns: among married groups, improvements in male academic performance are associated with significant reductions in dropout risk, while females show a flatter response curve. This may reflect the influence of traditional gender role expectations, where married females may face greater domestic responsibilities that can undermine academic persistence regardless of academic achievement.

I also found that academic progression (improvement from first to second semester) had differential impacts across groups: strongest for married male students and weakest for married female students. This suggests that academic trajectories over time are also moderated by gender and marital status.

These findings collectively support our central hypothesis that marital status significantly moderates the relationship between gender, academic performance, and dropout risk, revealing unique challenges and advantages faced by different demographic groups.

Discussion

The analysis provides compelling evidence that marital status moderates the relationship between gender and academic performance in predicting university student dropout. Several key findings emerge from this study. First, the conditional effect analysis demonstrates that marriage appears to have a protective effect against dropout for male students, particularly those with higher academic performance. In contrast, for female students, the protective effect of marriage is less pronounced and more variable across academic performance levels. The observed relationship between academic performance and dropout probability also displays different gradients across demographic groups. Male married students show a steeper decline in dropout risk as academic performance improves, with approximately a 43 percentage point decrease (from about 100% to 57%), suggesting they benefit more from academic achievement. Female students, however, exhibit more complex patterns that vary based on academic performance levels, with married female students showing the least responsiveness to academic performance improvements, experiencing only about a 63 percentage point decrease (from approximately 100% to 37%) in dropout probability.

Further analysis highlights resource allocation differences, with the Random Forest variable importance ranking indicating that financial strain indicators have higher predictive importance for specific demographic groups. This effect is particularly pronounced for married female students, where financial strain ranks significantly higher as a predictor of dropout risk than for other groups. This finding aligns with Kimmel et al. (2012), who suggested that differential role conflicts based on gender may contribute to variations in academic persistence. The protective effect of marriage for male students may also reflect additional support systems or motivational factors. As Carney-Crompton and Tan (2002) proposed, marriage can provide structure and purpose, but our findings indicate that this benefit may be gender-differentiated, potentially due to traditional gender roles and expectations.

Another key insight is performance consistency, where an analysis of first and second-semester academic metrics reveals that married female students exhibit the highest variability between semesters. This suggests differing patterns of adaptation to academic challenges, potentially linked to fluctuating family responsibilities. These findings have significant implications for understanding student retention, as the intersection of gender and marital status creates distinct risk profiles that cannot be fully captured by examining either factor in isolation. The moderating effect of marital status underscores the necessity of considering intersectional identities when predicting and addressing dropout risks.

While our modeling approach leveraged multiple machine learning algorithms to capture complex interactions, the tuned Random Forest model underperformed compared to other models,

including LASSO regression and Support Vector Machines. Despite being optimized through hyperparameter tuning, the tuned Random Forest model exhibited the lowest accuracy (0.8178) and the weakest predictive performance across several key metrics, including Positive Predictive Value (PPV) and F1-score. This suggests that while Random Forest can be a powerful tool for capturing non-linear relationships, it may be less effective in this context compared to more structured approaches like LASSO, which provided the best balance of sensitivity and specificity. The high importance ranking of the gender $\times$ marital status interaction term in both LASSO and Support Vector Machine models further reinforces the predictive value of these demographic factors in dropout risk analysis.

Research Limitations

Several limitations of this study should be acknowledged. One key limitation is the binary gender classification used in the dataset, which restricts our ability to explore more nuanced gender identities and their interactions with marital status. While modern understandings of gender extend beyond a simple binary framework, the available data do not allow for such complexity in analysis. Additionally, although marital status is recorded, detailed information about family structure—such as whether participants have children or the duration of their marriage—is absent. These missing details could provide deeper insights into how marital dynamics influence academic outcomes, limiting the scope of our interpretations.

Another significant limitation is the use of cross-sectional data, which constrains our ability to establish causal relationships. Longitudinal data would be more effective in examining how changes in marital status impact academic trajectories over time, but this is beyond the scope of the current dataset. The study is also constrained by limited diversity metrics, as the dataset lacks detailed information on race, ethnicity, and sexuality. The absence of these intersectional factors restricts a more comprehensive analysis of how multiple identity dimensions shape student experiences and dropout risks.

Furthermore, sample size limitations pose challenges, particularly for specific subgroups. While the overall sample size ($n = 4{,}424$) is sufficient, the representation of married students ($n = 546$) is relatively low, potentially reducing statistical power for interaction effect analyses, especially when further stratified by academic performance. Finally, measurement limitations must be considered, as some critical variables, such as financial strain and academic progression, rely on proxy indicators that may not fully capture their complexity. Financial strain, for instance, extends beyond tuition and debt to include unmeasured aspects like cost of living and family economic responsibilities, which are not explicitly accounted for in the dataset.

Future Research Directions

Future research should address these limitations by expanding demographic and family structure data collection to include more nuanced gender identity measures, marriage duration,

number of children, and distribution of family responsibilities. Implementing longitudinal studies would allow for tracking changes in marital status and academic performance over time, providing stronger evidence for causal relationships and shedding light on the evolving dynamics of student persistence. Additionally, incorporating qualitative research methods, such as interviews and focus groups, could offer deeper insights into the mechanisms behind the observed patterns, particularly the unique challenges faced by married female students.

Another crucial direction involves exploring intervention strategies specifically designed to support married female students, who often face distinct barriers. These may include financial assistance programs, childcare support services, and flexible academic policies that accommodate the additional responsibilities of married students. Moreover, future research should work toward developing more sophisticated theoretical models that account for multiple intersecting identities, moving beyond simple interaction effects to more comprehensive frameworks that explain how various factors collectively shape academic trajectories. Finally, conducting cross-validation studies in diverse higher education settings—including different institution types, countries, and cultural contexts—would help assess the generalizability of these findings and refine understanding of the specific factors influencing student retention across different environments.

Conclusion

This study demonstrates that marital status significantly moderates the relationship between gender and academic performance in predicting university student dropout. The findings reveal complex interaction effects that challenge simplistic understandings of either gender or marital status as isolated predictors of academic persistence.

The protective effect of marriage for academically successful male students (reducing dropout risk from approximately 100% to 57% as academic performance increases from very low to high levels) contrasted with the lesser impact for married female students (whose dropout rates only decrease from about 100% to 37% across the same performance spectrum) highlights the need for targeted support systems that consider intersectional identities. Higher education institutions should consider developing tailored retention strategies that address the specific challenges faced by different demographic groups, particularly married female students who appear to face unique barriers to academic persistence even when academically successful.

By understanding these complex interactions, institutions can move beyond one-size-fits-all retention approaches toward more nuanced strategies that recognize the diverse experiences and needs of their student populations.

References

Carney-Crompton, Shawn, and Josephine Tan. 2002. "Support Systems, Psychological Functioning, and Academic Performance of Nontraditional Female Students." *Adult Education Quarterly* 52 (2): 140–54.

Kimmel, Sara B, Kristena P Gaylor, M Ray Grubbs, and J Bryan Hayes. 2012. "Good Times to Hard Times: An Examination of Adult Learners' Enrollment from 2004-2010." *Journal of Behavioral and Applied Management* 14 (1): 18.

Negy, Charles et al. 2003. "Undergraduate Students' Adaptation to College: Does Being Married Make a Difference?" *Journal of College Student Development* 44 (5): 670–90.

Severiens, Sabine, and Geert Ten Dam. 2012. "Leaving College: A Gender Comparison in Male and Female-Dominated Programs." *Research in Higher Education* 53: 453–70.

Stinebrickner, Ralph, and Todd Stinebrickner. 2014. "Academic Performance and College Dropout: Using Longitudinal Expectations Data to Estimate a Learning Model." *Journal of Labor Economics* 32 (3): 601–44.