Demographic Factors and College Completion

Saúl Martínez

Advanced Data Analytics, University of North Texas

ADTA 5410 Applications and Deployment of Advanced Analytics

Dr. Jamie Humphries

December 9, 2024

Demographic Factors and College Completion

Abstract

This research investigates the strongest demographic predictors of college degree completions using data from over 16,000 U.S. higher education institutions for the 2022–2023 academic year. The study employs logistic regression, decision trees, and random forests to analyze the relationship between degree completions and predictors such as gender, age, and race/ethnicity. Female completions and completions among non-traditional students aged 25–39 were identified as the most significant factors influencing institutional success, followed by male completions and White student completions. Random forests emerged as the most robust model, achieving a 98.6% mean accuracy in 5-fold cross-validation, outperforming decision trees and logistic regression. These findings highlight the critical role of non-traditional students in driving completions and underscore the need for targeted interventions to support diverse demographics. The research provides actionable insights for policymakers and educational institutions aiming to improve degree completion rates across all student populations.

Introduction

The successful completion of college degrees is critical in driving societal progress and mobility for all communities and populations in the United States of America. Understanding the demographic factors that influence degree completion rates is crucial for developing targeted interventions and support systems that can contribute to higher student success in this area. This data set from the National Center for Education Statistics examines data from over 16,000 higher education institutions and analyzes degree completion rates from 2022-2023 between student demographic characteristics. This research seeks to identify the strongest predictors based on these demographic characteristics of degree completion. The predictors include demographic characteristics such as race, gender and age.

Research Question:

What demographic factors are the strongest predictors of degree completion across different higher education institutions?

Objectives:

- To quantify the relative influence of various demographic factors (gender, race/ethnicity, and age) on degree completion rates
- To identify patterns of success across different demographic groups and institutional types
- To develop insights that can inform evidence-based policies and interventions for improving completion rates among diverse student populations.

Significance:

This research holds particular significance in the current higher education landscape for several reasons:

- Equity and Access: By understanding demographic patterns in degree completion, institutions can better address disparities and promote educational equity.
- Resource Allocation: Insights from this study can help institutions optimize their support services and interventions for different student populations.
- Policy Implementation: Findings can inform evidence-based policies at both institutional and national levels to improve completion rates among diverse student groups.

Literature Review

Article 1:

Flores, S. M., & Park, T. J. (2013). Race, Ethnicity, and College Success: Examining the Continued Significance of the Minority-Serving Institution. *Educational Researcher*, *42*(3), 115–128. http://www.jstor.org/stable/23462375

This article focuses specifically on the role of MSIs (minority-serving institutions) and attempts to address the growth of these institutions since the 1980s and what kind of impact they have had on college completion. It analyzed data from Texas's Minority-Serving Institutions (MSIs) between 1997 and 2008 to evaluate postsecondary enrollment and completion outcomes by race and ethnicity. Using state administrative data and logistic regression models, the authors investigated the relationship between precollege characteristics, institutional factors, and educational outcomes.

The study found a particularly important finding and that is that race was a strong predictor of college enrollment but that at the college completion stage, race was no longer much of a factor, which means that once enrolled, race did not show much of a difference in college completion compared to their white counterparts. Although it is comprehensive longitudinal data (1997-2008) and large-scale state data from the state of Texas, it is limited to that specific state and it is limited to only addressing the impact of factors addressing college completion in MSIs only.

The methodology is robust, integrating longitudinal data and advanced statistical models to provide nuanced insights. However, the study's reliance on state-specific administrative data may limit the generalizability of its findings. Furthermore, its inability to account for variables such as

SAT/ACT scores, parental education, and generational status is a notable limitation. This is definitely important research as it pertains to addressing a very specific demographic (race) in a very specific state (Texas) during a very specific time period (1997-2008). However, there is a lot of room for improvement since educational policy and opportunities have evolved since the study. Since that period there has been a growing importance in online education (both in live and remote settings) in colleges in Texas that may affect the results of college completion even with this particular and specific setting. Furthermore, this data set and research is far from addressing the more diverse and comprehensive set of demographic features that exist.

Article 2:

Heil, S., Reisel, L., & Attewell, P. (2014). College Selectivity and Degree Completion. *American Educational Research Journal*, *51*(5), 913–935. http://www.jstor.org/stable/24546730

This research attempts to address the question, “How much of a difference does it make whether a student of a given academic ability enters a more or a less selective four-year college?” According to the abstract the authors of the study utilized, “...multilevel models and propensity score matching methods to reduce selection bias…” and found that “...selectivity does not have an independent effect on graduation.”

The study initially found a significant relationship between college selectivity (measured by average SAT scores) and graduation rates. However, this effect diminished when individual and institutional variables (e.g., socioeconomic status, tuition) were included in the analysis. Tuition emerged as a stronger predictor of graduation rates than selectivity. A $1,000 increase in annual tuition was associated with a small but statistically significant increase in graduation probabilities. Contrary to theories suggesting that under qualified students struggle at selective colleges, the study found no significant negative impact of "over-matching" or "under-matching" on graduation rates.

Some gaps in this study include the fact that the research here does not explore or include the role of other social or specific demographic factors and how they can affect this concept of selectivity and its effect on graduation. The study also seems to solely rely on SAT scores. The tuition component and its link to graduation rates may need to have further exploration of its predictive power. Is it that these institutions invest in necessary resources that provide the much needed support that students need to be successful? Or is that the higher tuition implies that students attending are already coming from higher income and more affluent families where they may have a lot more support than others in institutions with lower tuition institutions who may not have this support.

Article 3:

Eller, C. C., & DiPrete, T. A. (2018). The Paradox of Persistence: Explaining the Black-White Gap in Bachelor's Degree Completion. *American Sociological Review*, *83*(6), 1171–1214. https://www.jstor.org/stable/48588589

This research also excels at addressing a very specific demographic (black students compared to white students) and the disparity that exists in the area of bachelor's degree completion. The authors of the study employed data from the Education Longitudinal Study of 2002 and the Integrated Postsecondary Education Data System. The research identifies pre-college resource discrepancies as the largest contributors to the gap. Prior to college, black students experience lower academic and socioeconomic resources that contribute to this gap, according to the authors' analysis. The study highlights a paradox: Black students, despite having fewer academic and socioeconomic resources, are more likely than similarly resourced White students to enroll in four-year colleges. This "paradoxical persistence" partially offsets the attainment gap while widening the completion gap due to higher dropout rates. Black students show higher likelihood of four-year college enrollment despite resource disadvantages, however, the college completion data is far from favorable. The authors seek to point out this discrepancy.

The study is focused specifically on black students' experience in traditional four-year colleges which seems to overlook other non-traditional educational paths such as community colleges and more recently, online learning paths. The study is focused on one particular factor and that is race and, this study, focuses primarily on Black students vs White students. It would be interesting to explore if this same "paradoxical persistence" among other demographic groups and also an exploration as to what other demographic or societal factors are impacting this gap.

Article 4:

Hillman, N. W., & Orians, E. L. (2013). Financial Aid's Role in Meeting State College Completion Goals. *Education Finance and Policy*, *8*(3), 349–363. https://www.jstor.org/stable/educfinapoli.8.3.349

The authors of this research explore how state financial aid policies influence college access and completion and offer recommendations for aligning financial aid with state educational attainment goals. The study highlights the importance of "high-tuition, high-aid" models and the potential negative impact of merit-based aid programs that disproportionately benefit higher-income students. While the analysis is grounded in empirical evidence, the study primarily synthesizes existing research rather than introducing new data. Its reliance on secondary sources limits its ability to account for regional and institutional variations in aid effectiveness.

This is an important area of focus since most discussion on financial aid centers around federal aid. State aid, however, can also make a significant impact by utilizing a financial aid system effectively that can contribute not only to student success but to the overall economy and progress of that state.

Although this area of study is important, it is definitely limited in its scope not necessarily because it is limited to the state level but because it seems to omit so many different non-financial factors like other demographics, academic advising opportunities, etc. It includes limited research on non-traditional and part-time students and other educational paths such as community colleges and online learning avenues.

Methodology, Results and Discussion

This study employs a quantitative research design using institutional-level data to examine demographic factors predicting degree completion across higher education institutions. The analysis utilizes data from the National Center for Education Statistics (NCES) for the 2022-2023 academic year, comprising a sample of approximately 1,000 institutions from a total population of 16,065 higher education institutions. The dataset contains demographic variables (e.g., race, age, gender), institutional characteristics, and degree completion outcomes. The analysis will identify key predictors of degree completion using statistical and machine learning methods.

Statistical Methods and Analytical Tools:

Exploratory Data Analysis (EDA):

Tools: Python (Pandas, Matplotlib, Seaborn).

Purpose: To summarize data distributions, identify missing values, and visualize relationships between variables (e.g., correlation heatmaps, scatter plots).

Regression Analysis:

Method: Logistic Regression.

Purpose: To estimate the probability of degree completion based on demographic and institutional characteristics.

Why: Logistic regression is appropriate for binary outcomes, such as degree completion (Yes/No).

Classification Models:

Methods: Decision Trees and Random Forests.

Purpose: To classify students into groups based on the likelihood of degree completion and rank the importance of predictors.

Why: These models handle non-linear relationships and interactions effectively, providing interpretable results.

Cross-Validation:

Purpose: To validate the accuracy and reliability of models.

Why: Ensures that the model performs well on unseen data.

Variables

The dependent variable is total degree completion (CSTOTL), representing the total number of students receiving awards/degrees at each institution during the period July 1, 2022, to June 30, 2023.

1. **Target Variable (Dependent):**
   - degree_completion: Binary indicator (1 = Degree Completed, 0 = Not Completed).
2. **Input Variables (Independent):**

- **Demographic Characteristics:**
  - age: Age at enrollment.
  - gender: Male/Female.
  - race_ethnicity: Categorical variable.
- **Institutional Characteristics:**
  - institution_type: Public/Private.
  - tuition_cost: Numeric variable.
  - student_faculty_ratio: Numeric variable.
- **Other Predictors:**
  - marital_status: Binary (Married/Not Married).
  - parent_status: Binary (Parent/Not Parent).

Steps for Data Analysis

1. **Preprocessing:**
   - Check for missing values and handle them appropriately (e.g., imputation or exclusion).

The data set includes extra columns labeled with an X ahead of the name. For example for the column labeled CSTOTLM (Number of men receiving degrees) there is an XCSTOTLM column that references the imputation code and process that was used for that column.

CSTOTLT	XCSTOTLM
511	R
217	R
3	R
56	R

Code values for item imputation variables Xvarname	
CodeValue	ValueLabel
A	Not applicable
B	Institution left item blank
C	Analyst corrected reported value
D	Do not know
G	Data generated from other data values
H	Value not derived - data not usable
J	Logical imputation
K	Ratio adjustment
L	Imputed using the Group Median procedure
N	Imputed using Nearest Neighbor procedure
P	Imputed using Carry Forward procedure
R	Reported
Z	Implied zero;

Although this is important information, for the purposes of our analysis, those columns will not be utilized, so, in the data cleaning process, the step was taken to remove those columns.

The following two steps (normalize and encode) were performed after the EDA process.

- Normalize continuous variables (e.g., tuition_cost, student_faculty_ratio) to ensure comparability.
- Encode categorical variables using one-hot encoding (e.g., gender, race_ethnicity).

2. **Exploratory Data Analysis:**
    - Generate descriptive statistics for each variable.
    - Create visualizations (e.g., histograms, boxplots) to understand distributions.
    - Identify correlations between input variables and the target variable.

The following descriptive statistics of the data were generated to begin exploring the data and basic patterns and relationships in the data.

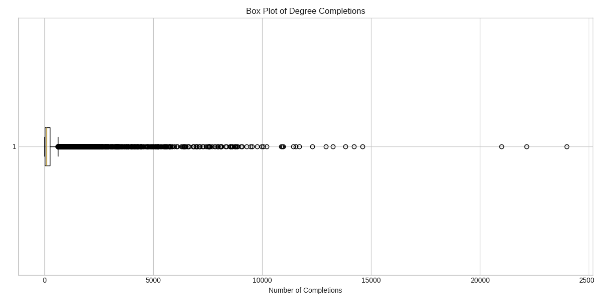
Degree Completion Distribution

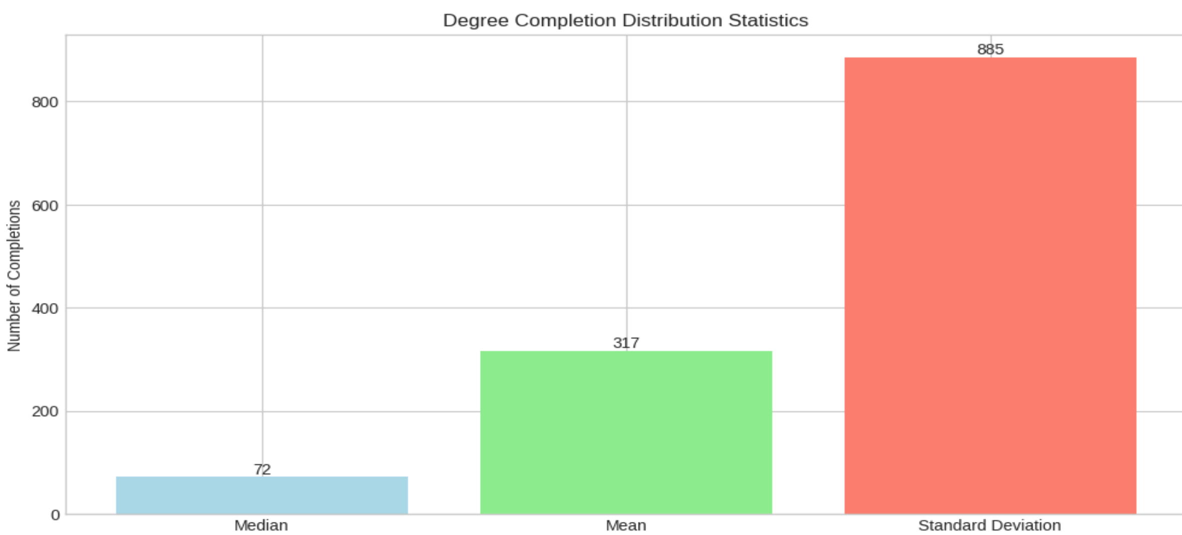
Mean completions per institution: 317

Median (50%): 72

Large spread (SD: 885) indicates high variability

Highly skewed right (mean > median)

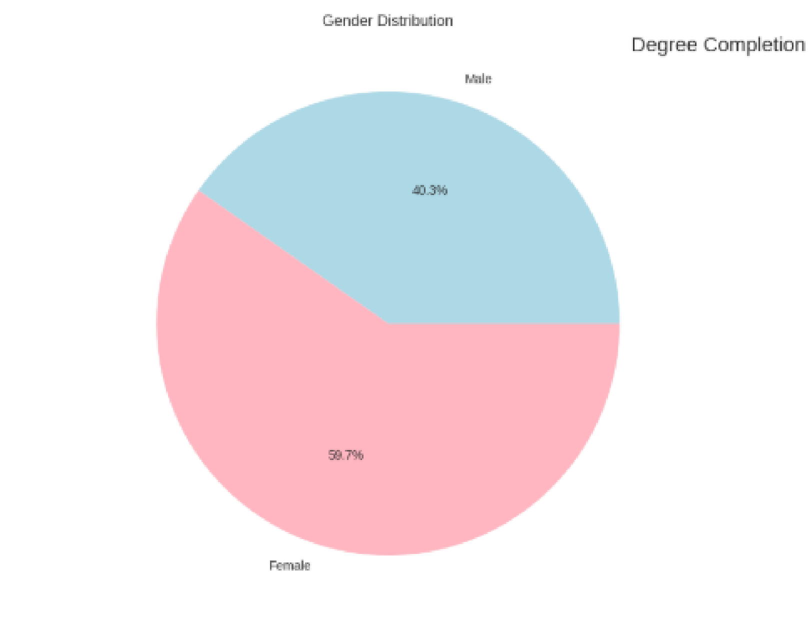




Gender Distribution

Female majority: 59.7%

Male: 40.3%



Race/Ethnicity Distribution

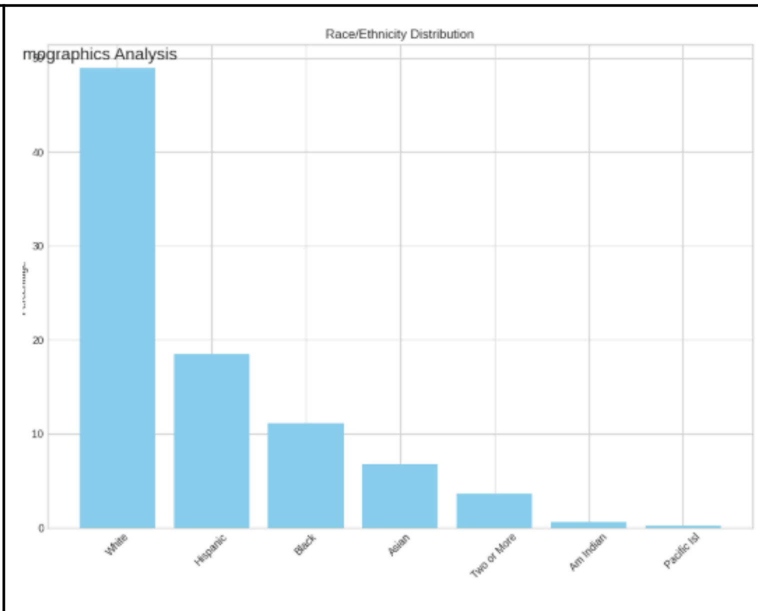
White students: 48.9% (plurality)

Hispanic: 18.5%

Black/African American: 11.2%

Asian: 6.8%

Other groups each < 4%



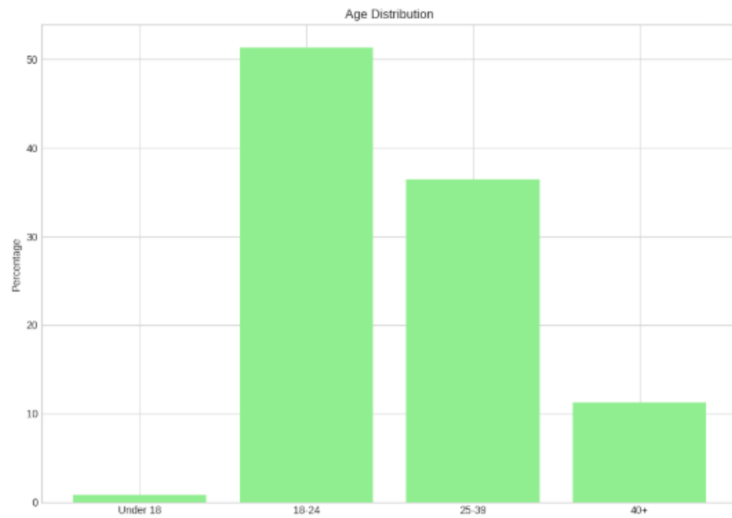
Age Distribution

Traditional college age (18-24): 51.3%

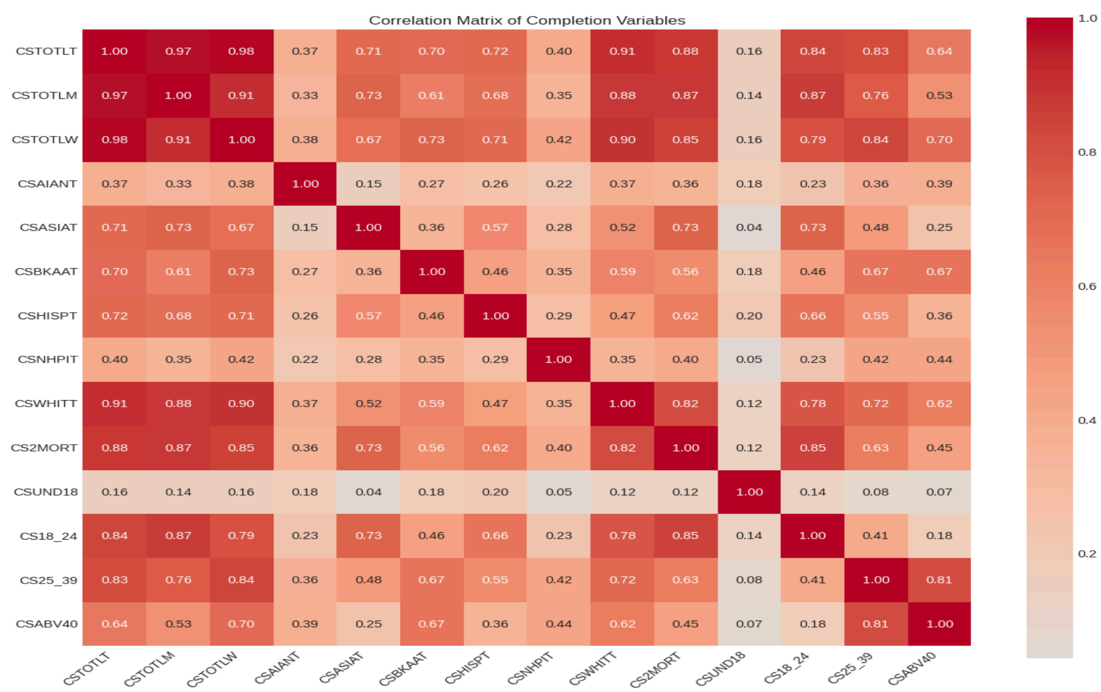
Adult learners (25-39): 36.5%

Older adults (40+): 11.3%

Under 18: < 1%



Correlation of Completion Variables:



1. Strong correlations (>0.9):

- Total completions (CSTOTLT) with male/female completions (CSTOTLM/CSTOTLW)
- White student completions (CSWHITT) with total completions (0.91)

2. Moderate to strong correlations (0.6-0.8):

- Age 18-24 completions with total completions (0.84)
- Age 25-39 with Age 40+ completions (0.81)

3. Weak correlations (<0.4):

- Under 18 completions (CSUND18) with most variables
- Native Hawaiian/Pacific Islander completions (CSNHPIT) with other variables

These patterns suggest degree completion rates tend to move together across demographic groups, with strongest relationships between total completions and major demographic categories.

Model Building:

- **Logistic Regression:** Fit the model to identify significant predictors.

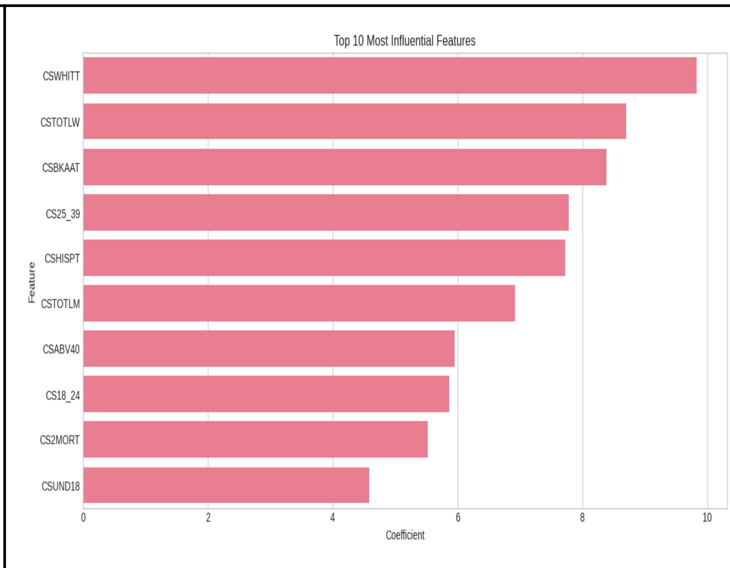
```
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.98      0.97      2365
           1       0.98      0.95      0.97      2455

    accuracy                           0.97      4820
   macro avg       0.97      0.97      0.97      4820
weighted avg       0.97      0.97      0.97      4820

Accuracy: 0.9651452282157676
```

The results of the logistic regression model showed an accuracy of 97% which is a strong indicator of model performance. The model had 95% precision on Class 0 (“Low Completions”) which means that when the model predicts “Low Completions” it is correct 95% of the time. Similarly, the model had even a higher precision (98%) on “High Completions”. In Recall, the model correctly identified 98% of the actual “Low completions” and 95% of the actual “High completions.” The F1-score combines precision and recall into a single metric and both classes had a high score of 97% suggesting a balance between precision and recall. The model seemed slightly better at identifying “low completions” (98%) than “high completions” (95%). The high accuracy and balanced precision/recall suggests that the model is effective at distinguishing between institutions with high vs. low completions. Logistic regression allows for identifying significant predictors by examining coefficients and we examined those next.

Coefficient Analysis

Feature Importance Analysis:

	Feature	Coefficient
8	CSWHITT	9.828559
2	CSTOTLW	8.703715
5	CSBKAAT	8.387998
14	CS25_39	7.784160
6	CSHISPT	7.727609
1	CSTOTLM	6.916899
15	CSABV40	5.951477
13	CS18_24	5.872082
9	CS2MORT	5.527913
12	CSUND18	4.589472



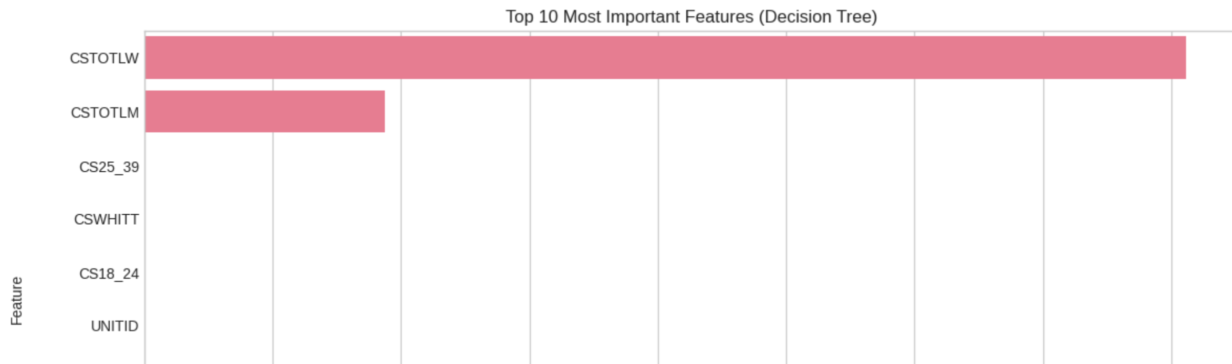
Positive coefficients (CWHITT, CSTOTLW) indicate that higher values of these features increase the likelihood of “High Completions" and vice versa. The White students’ completions (9.83) indicate that they have the strongest positive influence on predicting “High completions” which means that institutions with higher numbers of White students' completions are significantly more likely to fall into the “High Completion” category. Female students (CSTOTLW) were next with a slightly less positive influence at 8.70 followed by Black students’ completion at an even slightly lower score of 8.39. In the age categories, older student groups (CS25_39, CSABV40) rank higher than traditional college-age students (CS18_24). This could indicate the growing importance of non-traditional students in driving completions. As far as gender, Female completions (CSTOTLW) are slightly more predictive of “High Completions” than male completions (CSTOTLM), which may reflect broader trends in higher education success by gender.

Decision Trees: Train models to evaluate predictor importance and capture non-linear relationships.

- Tune hyperparameters (e.g., tree depth, number of estimators) for optimal performance.

```
Decision Tree Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      2365
           1       1.00      0.99      0.99      2455

    accuracy                           0.99      4820
   macro avg       0.99      0.99      0.99      4820
weighted avg       0.99      0.99      0.99      4820

Decision Tree Accuracy: 0.9925311203319502
```

```
Top 10 Feature Importances
CSTOTLW           0.806768
CSTOTLM           0.189932
CS18_24           0.001441
UNITID            0.000776
CS25_39           0.000533
CSWHITT           0.000521
CSHISPT           0.000029
CSUNKN            0.000000
Award_Level_11    0.000000
Award_Level_10    0.000000
dtype: float64
```



Pre-Tuning Results

Accuracy: 99.25% (before tuning).

- This is an extremely high score, suggesting the model already performs well without parameter optimization.

Precision, Recall, F1-Score:

- Both classes (0: Low Completions, 1: High Completions) have nearly perfect precision and recall, meaning the model is highly effective at distinguishing between these two classes.
- F1-scores for both classes are 0.99, indicating a strong balance between precision and recall.

Post-Tuning Results

1. **Tuned Parameters**:
    - **max_depth=20**: The optimal depth of the tree is 20, meaning the model can split the data more deeply to capture finer patterns.

- **min_samples_leaf=2**: Each leaf node must contain at least 2 samples, reducing the risk of overfitting.
- **min_samples_split=2**: Splits are allowed with at least 2 samples, keeping splits flexible.

2. **Accuracy**: 99.63% (after tuning).
   - A slight improvement over the pre-tuning accuracy, indicating that the optimized parameters marginally enhanced the model's performance.
3. **Precision, Recall, F1-Score**:
   - Scores remain consistent with pre-tuning results, further confirming the model's strong performance.

Observations and Insights

1. **High Accuracy Across Both Models**:
   - Both pre-tuned and tuned decision trees demonstrate excellent accuracy. This suggests the dataset is well-structured, and the predictors are highly informative.
2. **Marginal Gain After Tuning**:
   - The improvement from tuning (accuracy increase from 99.25% to 99.63%) is minor because the default model was already well-calibrated.
   - However, tuning helps ensure the model generalizes well and isn't overfitting.
3. **Balanced Class Performance**:
   - Both classes (Low and High Completions) are predicted with equal strength, showing no bias in the model.

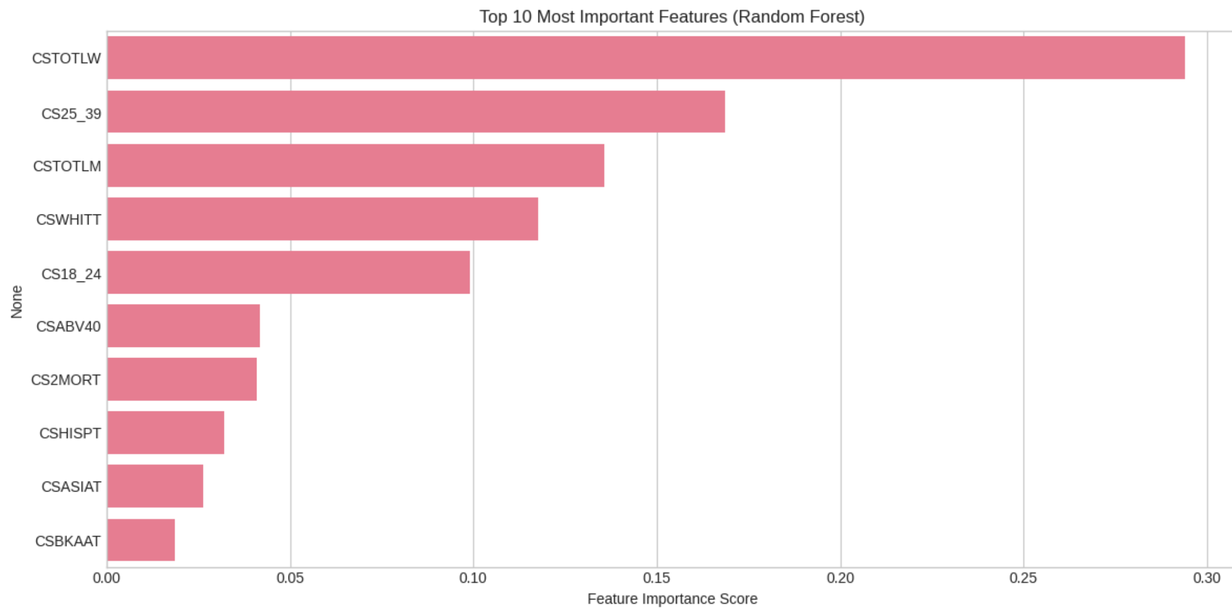
The tree's success highlights that certain predictors (e.g., completions by gender, race, and age groups) are highly discriminative. For instance:

- Predictors like CSTOTLW (Female completions) and CSWHITT (White student completions) likely create early splits in the tree, separating institutions with high completions from those with low completions.
- Age group completions (CS18_24, CS25_39) likely refine the classification, capturing the contribution of non-traditional students to institutional success.

Random Forest: Train models to evaluate predictor importance and capture non-linear relationships.

- ○ Tune hyperparameters (e.g., tree depth, number of estimators) for optimal performance.

```
Best Parameters: {'max_depth': 15, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 200}
Accuracy for Random Forest: 0.9937759336099585
Classification Report for Random Forest:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      2365
           1       0.99      0.99      0.99      2455

    accuracy                           0.99      4820
   macro avg       0.99      0.99      0.99      4820
weighted avg       0.99      0.99      0.99      4820
```



Random Forest Results

1. **Best Parameters**:
    - max_depth=15: The trees in the forest have a maximum depth of 15, allowing them to capture meaningful complexity without overfitting.
    - min_samples_leaf=2: Each leaf node contains at least two samples, ensuring meaningful splits.
    - min_samples_split=5: Nodes split only if they have at least five samples, adding robustness.
    - n_estimators=200: The forest aggregates predictions from 200 trees, improving stability and reducing overfitting.

Accuracy:

- **Test Accuracy**: 99.38%

    - Slightly higher than the decision tree (99.25% pre-tuning, 99.63% post-tuning) and comparable to logistic regression (96.5%).
- **Implication**: The random forest benefits from ensemble learning, combining multiple trees to reduce bias and variance, leading to robust predictions.

Classification Report:

- Both classes (Low Completions = 0, High Completions = 1) have:
    - Precision, Recall, and F1-Score of 0.99.
    - This confirms that the model performs equally well in identifying both "High" and "Low Completions" without bias.
- **Implication**: The random forest maintains high predictive power for both classes, similar to the decision tree.

Feature Importance:

- **Top Features**:
    - CSTOTLW (Female Completions): The most important predictor, contributing nearly 30% to the model's decisions.
    - CS25_39 (Age 25–39 Completions): Highlights the importance of non-traditional students in completions.
    - CSTOTLM (Male Completions): Male completions also strongly predict total completions.
    - CSWHITT (White Completions): Indicates the contribution of White students to institutional success.

  - CS18_24 (Age 18–24 Completions): Traditional college-age students remain key contributors but rank below non-traditional age groups.
- **Lower Ranked Features**:
  - CSASIAT (Asian Completions), CSBKATT (Black Completions), and CSHISPT (Hispanic Completions) rank lower, reflecting their relatively smaller total completion numbers in the dataset.
- **Implication**: The results align with decision tree and logistic regression findings, where demographic breakdowns (gender, age, race) drive completions, with slight differences in feature ranking.

Comparison with Decision Tree and Logistic Regression

1. Accuracy:

- **Random Forest**: 99.38% (best model due to ensemble learning).
- **Decision Tree**: 99.63% (tuned), but this model may slightly overfit due to a single-tree structure.
- **Logistic Regression**: 96.5% (solid performance but slightly lower due to its linear nature).
- **Key Insight**: Random forests combine multiple trees to generalize better and avoid overfitting compared to decision trees.

2. Feature Importance:

- **Random Forest**:

  - CSTOTLW (Female Completions) is the most important, followed by age groups and CSTOTLM (Male Completions).
- **Decision Tree**:
  - CSTOTLW also ranked highest, followed by CSWHITT (White Completions) and CS25_39 (Age 25–39).
- **Logistic Regression**:
  - CSWHITT (White Completions) was the strongest predictor, followed by CSTOTLW (Female Completions).
- **Key Insight**:
  - Random forests and decision trees agree on the importance of gender (CSTOTLW, CSTOTLM) and age group completions (CS25_39).
  - Logistic regression emphasizes racial demographics (e.g., CSWHITT).

3. Flexibility:

- **Random Forest**: Handles non-linear relationships well and is robust against outliers due to ensemble aggregation.
- **Decision Tree**: Captures non-linear patterns but is prone to overfitting, especially at greater depths.
- **Logistic Regression**: Simple and interpretable but limited to linear relationships.

Key Insights from Random Forest

1. **Non-Traditional Students Matter**:

- The importance of CS25_39 and CSABV40 underscores the growing role of non-traditional students in degree completions.
- Institutions might benefit from tailoring support services for older age groups.

2. **Gender Balance**:
   - Both female (CSTOTLW) and male completions (CSTOTLM) are critical, with female completions having a slightly stronger impact.
3. **Broad Demographic Patterns**:
   - While completions by White students (CSWHITT) dominate the racial category, completions by other racial groups (e.g., CSBKATT, CSHISPT) also contribute, albeit to a lesser degree.
4. **Validation Across Models**:
   - The consistent ranking of key features (e.g., CSTOTLW, CSTOTLM, CSWHITT) across logistic regression, decision trees, and random forests validates their importance.

Model Validation:

- Use k-fold cross-validation to evaluate the accuracy and generalizability of models.
- Compare models using metrics such as accuracy, precision, recall, and F1 score.

K-fold cross validation helps to ensure the results observed in the decision tree and the random forest will generalize well across different subsets of data. A 5-fold cross validation is performed on the decision tree and random forest.

Decision Tree:

```
Decision Tree Cross-Validation Scores (5-fold): [0.9956427  0.99657641 0.99595394 0.99470899 0.5664488 ]
Decision Tree Mean Accuracy: 0.9099
```

The first four folds show high accuracy (above 99%) suggesting that the model performs well on certain sets of the data. However, the last set dropped significantly to 57% suggesting that the model may be overfitting to certain sets of the data.

Random Forest:

```
Random Forest Cross-Validation Scores (5-fold): [0.98817305 0.99346405 0.99502023 0.98941799 0.96389667]
Random Forest Mean Accuracy: 0.9860
```

The random forest shows a more consistent pattern with high scores across all five folds keeping the mean accuracy at 99% which suggests a strong ability to generalize across all subsets of the data.

Unlike the decision tree, the random forest’s scores are much less variable and shows it's a better model at handling different data patterns more effectively. Random forests aggregate predictions from multiple decision trees, reducing overfitting and variance. This aggregation makes the model more robust. The random forest is the better model over the single decision tree achieving high accuracy and low variability.

Conclusion

This study analyzed the demographic factors influencing college degree completion rates across over 16,000 higher education institutions. Key findings indicate that gender, age, and race/ethnicity are significant predictors of degree completions. Female completions (CSTOTLW) emerged as the most critical factor, followed closely by male completions (CSTOTLM) and completions by non-traditional students aged 25–39 (CS25_39). Completions among White students (CSWHITT) also strongly contributed, while completions among Black (CSBKATT) and Hispanic students (CSHISPT) had a smaller but notable influence.

Logistic regression, decision trees, and random forests provided complementary insights. Logistic regression highlighted linear relationships, emphasizing racial factors, while decision trees and random forests excelled at capturing non-linear interactions and ranking predictors. Random forests proved the most robust and generalizable, achieving an accuracy of 98.6% during 5-fold cross-validation, outperforming the decision tree and logistic regression models.

The findings underscore the growing importance of non-traditional students in driving degree completions, highlighting opportunities for institutions to tailor interventions to support diverse demographics. Future research should explore additional factors, such as socioeconomic background and institutional characteristics, to develop a more comprehensive understanding of degree completion patterns. Additionally, longitudinal studies could examine how changing demographic trends impact degree completions over time.

References

Eller, C. C., & DiPrete, T. A. (2018). The Paradox of Persistence: Explaining the Black-White Gap in Bachelor's Degree Completion. *American Sociological Review*, *83*(6), 1171–1214. https://www.jstor.org/stable/48588589

Flores, S. M., & Park, T. J. (2013). Race, Ethnicity, and College Success: Examining the Continued Significance of the Minority-Serving Institution. *Educational Researcher*, *42*(3), 115–128. http://www.jstor.org/stable/23462375

Heil, S., Reisel, L., & Attewell, P. (2014). College Selectivity and Degree Completion. *American Educational Research Journal*, *51*(5), 913–935. http://www.jstor.org/stable/24546730

Hillman, N. W., & Orians, E. L. (2013). Financial Aid's Role in Meeting State College Completion Goals. *Education Finance and Policy*, *8*(3), 349–363. https://www.jstor.org/stable/educfinapoli.8.3.34