Demographic Factors and College Completion

An Analysis of U.S. Higher Education Institutions

Saúl Martínez

ADTA 5410 : Applications and Deployment of Advanced Analytics, Fall 2024

THE DATA SOURCE

Integrated Postsecondary Education Data System

IPEDS is a system of interrelated surveys conducted annually by the US Department of Education's National Center for Education Statistics (NCES).

nces.ed.gov

THE DATA SET

16,065 U.S. Higher Education Institutions

The data set contains ***the number of students receiving a degree or certificate*** by level of award and by ***race/ethnicity, gender and age*** categories.

The data set covers awards granted between ***July 1, 2022 and June 30, 2023.***

INTRODUCTION

What demographic factors are the strongest predictors of degree completion?

Race/Ethnicity, Gender and Age

LITERATURE REVIEW

A Brief Overview

Literature Review

- Focus: Minority-Serving Institutions (MSIs) in Texas (1997-2008)
- Findings: Race strongly predicts enrollment, but not completion
- Strengths: Longitudinal data, robust methodology
- Limitations: State-specific data, lacks variables like SAT scores

Article 1

Race, Ethnicity, and College Success:

Examining the Continued Significance of

the Minority-Serving Institutions

Flores, S. M., & Park, T. J. (2013).

Literature Review

- Focus: College selectivity and degree completion
- Findings: Tuition predicts graduation more than selectivity
- Strengths: Addresses ‘overmatching’ and ‘undermatching’ theories.
- Limitations: Limited exploration of demographics, reliance on SAT scores

Article 2

College Selectivity and Degree Completion.

Heil, S., Reisel, L., & Attewell, P. (2014).

Literature Review

- Focus: Black-White gap in bachelor's degree completion
- Findings: Resource disparities drive gaps; paradoxical persistence
- Strengths: Highlights pre-college resource discrepancies
- Limitations: Focuses on traditional 4-year colleges, overlooks other paths

Article 3

The Paradox of Persistence: Explaining the Black-White Gap in Bachelor's Degree Completion.

Eller, C. C., & DiPrete, T. A. (2018).

Literature Review

- Focus: State financial aid policies and college completion
- Findings: High-tuition, high-aid models improve outcomes
- Strengths: Highlights state aid's role in educational goals
- Limitations: Omits non-financial factors and non-traditional students

Article 4

Financial Aid's Role in Meeting State College Completion Goals

Hillman, N. W., & Orians, E. L. (2013).

INTRODUCTION

What demographic factors are the strongest predictors of degree completion?

Race/Ethnicity, Gender and Age

OBJECTIVES

- To quantify the influence of demographic factors (gender, race/ethnicity, age).
- To identify patterns of success across demographic groups
- To develop insights that can Inform evidence-based policies for improving completions rates.

SIGNIFICANCE

- **Address educational equity**
- Guide resource allocation
- Inform educational policies

THE METHODOLOGY

The Variables

Dependent: Degree completions (CSTOTL)

Independent: Gender (CSTOTLW, CSTOTLM) age (Various), race/ethnicity (Various)

THE METHODOLOGY

Machine Learning Models

MODELS:

Logistic Regression: Linear Relationships

Decision Trees : Non-linear Relationships

Random Forest : Aggregate predictions from multiple decision trees, reducing overfitting and improving accuracy

Validation : 5-fold cross validation

THE ANALYSIS

Logistic Regression

Linear Relationship

Estimate the probability of degree completion based on demographic characteristics

Accuracy : 97%

Key Predictors : CSWHITT (White completions) and CSTOTLW (Female completions)

THE ANALYSIS

Logistic Regression

```
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
```

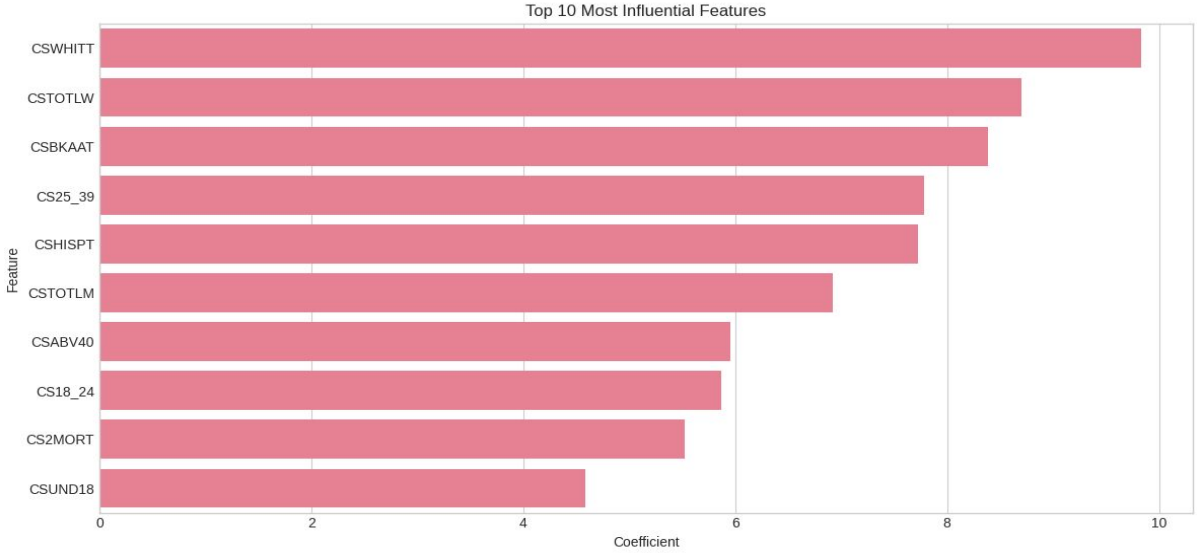
THE ANALYSIS

Logistic Regression

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



THE ANALYSIS

Decision Tree

Non-Linear Relationship

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

Accuracy : 99.25 (pre-tuned) 99.63% (tuned)

Key Predictors : CSTOTLW (Female completions), CSTOTLM (Male), CSWHITT (White)

THE ANALYSIS

Decision Tree

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

THE ANALYSIS

Random Forest

Multiple Decision Trees

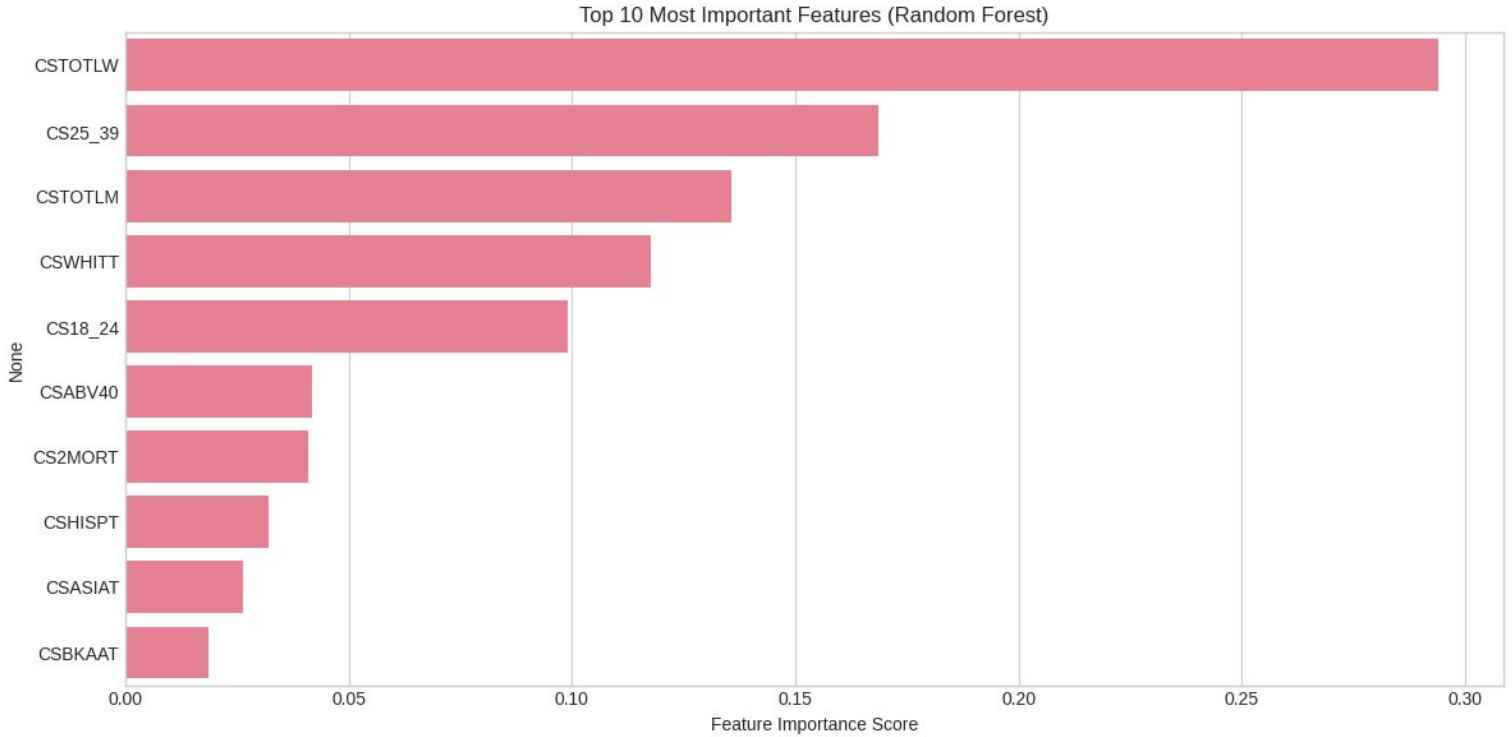
To evaluate predictor importance and capture non-linear relationships more robustly than a single decision tree.

Accuracy : 99.38% (consistent across folds)

Key Predictors : CSTOTLW (Female completions), CSWHITT (White completions) and CS25_39 (Age - Non-traditional students)

THE ANALYSIS

Random Forest



THE ANALYSIS

Model Validation

K-fold Cross-Validation

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

THE ANALYSIS

Model Validation

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
```

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
```

CONCLUSION

Key Takeaways

Female completions were the strongest predictor.

Non-traditional students (25_39) play a critical role.

Random Forest demonstrated superior accuracy and reliability.

CONCLUSION

Recommendations

Provide tailored interventions to support diverse demographics, particularly, non-traditional students.

Future research would be needed to expand on factors missing in this study such as socioeconomic and institutional factors.

These factors would help provide a more comprehensive understanding of college completion rates.