Gender and Higher Education: An Analysis of IPEDS Data

Cori N. Miller

Northwest Missouri State University, Maryville MO 64468, USA
s545767@nwmissouri.edu

Abstract. Keywords: data analytics · higher education

1 Introduction

Gender equity in higher education is a critical issue with far-reaching implications for individuals and society as a whole. Despite significant progress in recent decades, gender gaps persist in higher education enrollment, persistence, graduation, and academic performance. This research paper will explore gender equity gaps in higher education and use gender to predict higher education outcomes.

The research will use a mixed-methods approach, combining quantitative data from The National Center for Education Statistics (NCES) and Integrated Post secondary Education Data System (IPEDS) with Machine learning models. These models will be developed to predict higher education outcomes using gender as a predictor variable.

This research has one key limitation: it focuses on gender as the only predictor variable. While gender is an important factor in higher education outcomes, it is not the only factor. Other factors, such as race, ethnicity, socioeconomic status, and first-generation status, also play a role. Future research should explore the intersection of gender with other factors to better understand the gender gaps in higher education.

1.1 Goals of this Research

Higher education is essential for individual and societal success. However, gender equity gaps persist in higher education, with women underrepresented in certain fields of study and at higher levels of education. This research will explore gender equity gaps in higher education and use gender to predict higher education outcomes. The independent variable is the percentage of the cohort who are identified as women, and the dependent variable is the overall completion rate. This research is expected to produce the following results:

- A more comprehensive understanding of the gender gaps in higher education
- Insights into how gender affects students' academic graduation outcomes
- Machine learning models that can be used to predict higher education outcomes using gender as a predictor variable
- Proposed interventions to help close the gender gaps in higher education

2 Data and Methods

The data for this study was collected from the National Center for Education Statistics (NCES) Graduation Rates Survey. A custom data set was created using the NCES data tools to focus on graduation rates by gender and population ratios. The data set includes 298 records and 14 fields of structured data in CSV format.

2.1 Data Preparation and Cleaning

This study uses a custom data set created from the National Center for Education Statistics (NCES) Graduation Rates Survey. The data set includes 298 records and 14 fields of structured data in CSV format, with a focus on graduation rates by gender and population ratios. The data is high quality, with no missing fields or formatting issues. The following steps were taken to collect and prepare the data:

- The Graduation Rates Survey was accessed through the NCES website.
- A custom list of comparison institutions was created based on their similarity to my own institution.
- The data was extracted from the Graduation Rates Survey for the specified institutions and variables.
- The data was inspected to identify any obvious errors or inconsistencies
- Missing values, duplicate rows, and invalid values were corrected
- Data outliers, including non-coed institutions, were removed
- Calculated Fields were added to

2.2 Data Attributes

The following table lists the data attributes that were used in this study:

- unitid: A unique identifier for the post secondary institution (Integer)
- institution name: The name of the post secondary institution (String)
- year: The year in which the data was collected (Date)
- FTFT Cohort Grand total: The total number of first-time, full-time (FTFT) students who enrolled in the post secondary institution in the fall of the year specified (Integer)
- FTFT Cohort Total men: The total number of FTFT male students who enrolled in the post secondary institution in the fall of the year specified (Integer)
- FTFT Cohort Total women: The total number of FTFT female students who enrolled in the post secondary institution in the fall of the year specified (Integer)
- Percent Men: The percentage of FTFT students who were male (Float)
- Percent Women: The percentage of FTFT students who were female (Float)

- Total Completers: The total number of students who completed their degree or certificate program at the post secondary institution within six years of their initial enrollment (Integer)
- Total Completers Men: The total number of male students who completed their degree or certificate program at the post secondary institution within six years of their initial enrollment (Integer)
- Total Completers Women: The total number of female students who completed their degree or certificate program at the post secondary institution within six years of their initial enrollment (Integer)
- Over All Grad Rate: The overall graduation rate for FTFT students, calculated as the percentage of FTFT students who completed their degree or certificate program within six years of their initial enrollment (Float)
- Grad Rate Men: The graduation rate for FTFT male students, calculated as the percentage of FTFT male students who completed their degree or certificate program within six years of their initial enrollment (Float)
- Grad Rate Women: The graduation rate for FTFT female students, calculated as the percentage of FTFT female students who completed their degree or certificate program within six years of their initial enrollment (Float)

2.3 Exploratory Data Analysis

Exploratory data analysis (EDA) is a crucial step in data analytics, involving investigating and summarizing a data set to comprehend its key features, identify patterns, and uncover any anomalies. It is akin to being a detective for your data, attempting to decipher the wealth of information it contains. EDA ensures a thorough understanding of the data before embarking on complex analysis.

For this project, a combination of descriptive statistics, visualizations, and limited regression analysis proved effective in conducting EDA. Python, along with the pandas and matplotlib packages, was employed for this purpose. The data's head and tail were examined, and statistical information was summarized. Included below are a histograms and correlation tables for key attributes.

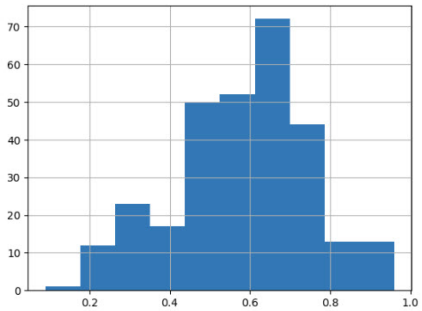

Fig. 1. Overall Graduation Rate Histogram

	PercentMen	PercentWomen	GradRate	GradRateMen	GradRateWomen
PercentMen	1.000000	-0.999927	0.199733	0.180961	0.259835
PercentWomen	-0.999927	1.000000	-0.200051	-0.181471	-0.260082
GradRate	0.199733	-0.200051	1.000000	0.966292	0.974815
GradRateMen	0.180961	-0.181471	0.966292	1.000000	0.891547
GradRateWomen	0.259835	-0.260082	0.974815	0.891547	1.000000

Fig. 2. Correlations of Key Attributes

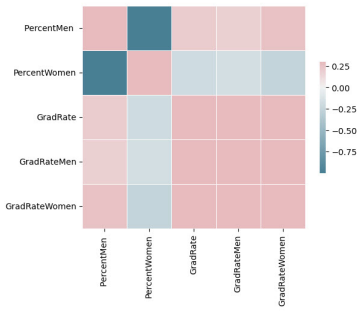

Fig. 3. Heatmap of Key Attributes

An assessment of data quality was conducted to ensure the validity of the data used in this study. This involved scrutinizing the data for missing values, outliers, and inconsistencies. The examination revealed that the data was complete, with no missing values or outliers identified. However, the correlation analysis revealed a relatively weak relationship between the percentage of women and overall completion rates. While the correlations were statistically significant, the magnitude of the relationships suggests that these variables may not be strongly predictive of one another. This finding warrants further investigation to determine the underlying factors influencing the observed relationships.

The Python notebooks and data used for this project are available at https://github.com/corimiller/CM_Capstone

3 Predictive Modeling

The predictive application utilized a number of predictive models to forecast student graduation rates based on the percentage of women in a cohort. This approach involved the following steps:

1. Data Collection and Preparation: Relevant student data was gathered from the National Center for Education Statistics. Data cleaning and preprocessing techniques were applied to ensure data consistency and quality.

2. Model Selection and Training: The scikit-learn library was employed to implement a linear regression model, decision tree model and a random forest model. (The Python notebooks and data used for this project are available at https://github.com/corimiller/CM_Capstone) These models were trained using a portion of the prepared data, allowing it to learn patterns and relationships between the input feature (percentage of women) and the target variable (graduation rate).
3. Model Evaluation and Tuning: The trained models were evaluated using the remaining portion of the data, assessing its ability to generalize to unseen data. Tuning techniques were applied to optimize the model's performance.
4. Model Interpretation and Application: The models were analyzed to understand the influence of the input feature on the target variable. The trained and optimized model was then applied to predict graduation rates for new cohorts

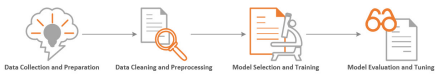

Fig. 4. Predictive Modeling Workflow

3.1 Modeling with Python

Below is a sample of the python code used to implement these models. The provided code snippet trains a linear regression model to predict student graduation rates based on the percentage of women in a cohort.

```
from sklearn.model_selection import train_test_split
train_set, test_set = train_test_split(auto,
 test_size=0.2, random_state=123)

print('Train size: ', len(train_set), 'Test size: ', len(test_set))

from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score

X = train_set[['PercentWomen']]
y = train_set['GradRate']

X_test = test_set[['PercentWomen']]
y_test = test_set['GradRate']
```

```
lr_model = LinearRegression()
lr_model.fit(X,y)

y_pred = lr_model.predict(X)
print('Results for linear regression on training data')
print('  Default settings')
print('Internal parameters:')
print('   Bias is ', lr_model.intercept_)
print('   Coefficients', lr_model.coef_)
print('   Score', lr_model.score(X,y))

print('MAE is  ', mean_absolute_error(y, y_pred))
print('RMSE is ', np.sqrt(mean_squared_error(y, y_pred)))
print('MSE is ', mean_squared_error(y, y_pred))
print('R^2    ', r2_score(y,y_pred))

y_test_pred = lr_model.predict(X_test)
print()
print('Results for linear regression on test data')

print('MAE is  ', mean_absolute_error(y_test, y_test_pred))
print('RMSE is ', np.sqrt(mean_squared_error(y_test, y_test_pred)))
print('MSE is ', mean_squared_error(y_test, y_test_pred))
print('R^2    ', r2_score(y_test,y_test_pred))
```

3.2 Linear Regression Model

Linear regression is a statistical method that assumes a linear relationship between the input feature (percentage of women) and the target variable (graduation rate). This assumption may not be valid for this data, as the relationship between the two variables may be more complex. Additionally, linear regression is not able to capture the nonlinear relationships and interactions between various factors that can impact graduation rates.

Metric	Training	Test
Bias	0.837455	0.102093
Coefficients	-0.470636	NA
Score	0.048751	NA
MAE	0.132752	0.131797
RSME	0.170724	0.173705
MSE	0.029147	0.029147
R Squared	0.048751	-0.153553

Table 1. Linear Regression Results.

The linear regression model has a lower mean squared error (MSE) on the test data than on the training data. This suggests that the model is generalizing well to unseen data. However, the R-squared score on the test data is negative, which means that the model is actually worse than random guessing at predicting the target variable.

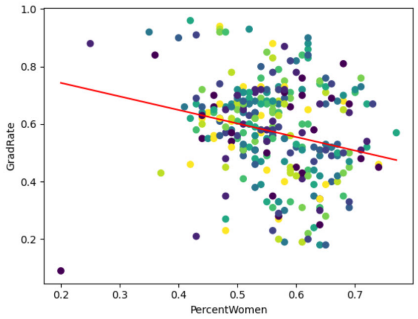

Fig. 5. Linear Regression of Graduation Rates

3.3 Decision Tree Model

Decision trees are a type of machine learning algorithm that can learn complex nonlinear relationships between input features and target variables. However, decision trees can be prone to overfitting, which means that they can learn the noise in the training data too well and not generalize well to unseen data.

The decision tree model achieved a mean squared error (MSE) of 0.022824 on the training data and an MSE of 0.019690 on the test data. This suggests that the model is generalizing well to unseen data. However, the R-squared score on the test data is -0.307593, which means that the model is not a reliable tool for predicting the target variable.

Metric	Training	Test
MSE	0.022824	0.019690
R Squared	0.255109	-0.307593

Table 2. Decision Tree Results.

3.4 Random Forest Model

Random forest is a type of ensemble machine learning algorithm that combines multiple decision trees. This approach can help to reduce the risk of overfitting, as

the individual decision trees can "vote" on the best prediction. Random forest is generally a good choice for predicting student graduation rates, as it can capture complex nonlinear relationships and is less prone to overfitting than decision trees.

The random forest model achieved a mean squared error (MSE) of 0.023774 on the training data and an MSE of 0.020319 on the test data. This suggests that the model is generalizing well to unseen data. However, the R-squared score on the test data is -0.349344, which means that the model performs worse than simply guessing the target variable.

Metric	Training	Test
MSE	0.023774	0.020319
R Squared	0.224106	-0.349344

Table 3. Random Forest Results.

4 Conclusion

Summary of the key findings

[]

References

1. https://nces.ed.gov/ipeds/
2. https://nces.ed.gov/
3. O'Connor, P.: Why is it so difficult to reduce gender inequality in male-dominated higher educational organizations? a feminist institutional perspective. Interdisciplinary Science Reviews (2020)