Enrollment Projections Using Machine Learning Algorithms

Meghaan L. Binkley-Hopper

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

Abstract. This project analyzes enrollment trends for public 4-year universities in the state of Missouri and seeks to make enrollment projections for Northwest Missouri State University. The data analysis for this project was conducted in 2 phases, exploratory data analysis and machine learning models. The exploratory data analysis was completed through descriptive statistics in Python and various visualizations. The results showed that enrollment for first-time freshmen in the state of Missouri is on a downward trend from 2012-2022. Linear regression, polynomial regression, and elastic net with polynomial features were the machine learning models applied to the Northwest data. The machine learning models did not produce useful correlations.

Keywords: data analytics · university enrollment · forecasting · Missouri · Northwest Missouri State University · higher education · linear regression · machine learning

1 Introduction

Higher Education Institutions are experiencing issues with enrolling traditional, 18 to 24 years old, students due to a shrinking population of high school graduates. [4] According to the Center for Disease Control (CDC), the birth rate for live births from 2009 to 2019 declined from 13.5 live births per 1000 to 11.4 live births per 1000 in the United States. [1] This year, colleges and universities in the United States experienced a 3.6 percent decline in freshman enrollment. [5] Due to the declining population of traditional students on college campuses, it is imperative to understand the trend, set reasonable expectations for the future, and plan innovative ways to make up the difference.

For the purpose of this research project, public 4-year institutions in the state of Missouri were analyzed. Data was collected from the Integrated Postsecondary Education Data System (IPEDS) and Missouri's Department of Elementary and Secondary Education (DESE). The IPEDS data contains enrollment statistics for public 4-year higher education institutions in the state of Missouri, and the DESE data contains enrollment statistics for kindergarten through the 12th grades in Missouri. Both data sources contain current and historical data.

1.1 Goals of this Project

There are 2 goals for this project. The first goal was to examine and compare enrollment trends for public 4-year institutions in the state of Missouri. This was accomplished through completing an exploratory data analysis with Python, Excel, and Tableau. The second goal was to examine the enrollment trends for Northwest Missouri State University and use them to train a machine learning model in Python to forecast future enrollment projections.

In order to accomplish these goals, the following implementation process was followed. (1)Data Collection, (2)Data Curation, (3)Exploratory Data Analysis Performance, (4)Machine Learning Performance, (5)Make Conclusions.

1.2 Project Links

This project is hosted on Overleaf, Github, and Tableau Public. Data files were sourced from IPEDS and DESE.

1. Overleaf Link: https://tinyurl.com/Binkley-HopperOverleaf
2. Github Link: https://tinyurl.com/Binkley-HopperGithubRepo
3. Tableau Public Link: https://tinyurl.com/Binkley-HopperTableau
4. IPEDS Link: https://tinyurl.com/Binkley-HopperIPEDSData
5. DESE Link: https://tinyurl.com/Binkley-HopperDESEData

1.3 Related Works

Various studies have been completed analyzing declining enrollment in higher education. One study by Benjamin Fields and Steven Brint looks at enrollment trends across public higher education institutions in the United States. This study looks at 3 different types of institutions, community colleges, regional 4-year institutions, and research institutions. Fields and Brint applied different lenses, such as political and economic, to each of these different types of institutions to determine whether or not there are any correlations to be made. [2]

In another article by Anthony Schuette, the idea of the "enrollment cliff" is discussed. The author posits that there was a decline in birth rate in the United States that began between 2007 and 2009. Schuette mentions that there was an economic recession during this time, but the birth rate did not recover when the economy rebounded. Furthermore, the percentage of students that go straight from high school to college has stayed consistent in recent years showing that higher education institutions are unlikely to be able to make up the difference by recruiting a larger percentage of graduates. Due to these issues, the author believes that higher education institutions will start to see a decline in enrollment starting in 2025. [7]

2 Data Curation

Data for this project was gathered from the Integrated Postsecondary Education Data System (IPEDS) and the Missouri Department for Elementary and Secondary Education (DESE) websites. The IPEDS data includes enrollment numbers of first-time freshmen from public, 4-year institutions from the state of Missouri. The universities that are included are Harris-Stowe State University, Lincoln University, Missouri Southern State University, Missouri State University, Missouri University of Science and Technology, Missouri Western State University, Northwest Missouri State University, Southeast Missouri State University, Truman State University, University of Central Missouri, University of Missouri-Columbia, University of Missouri-Kansas City, and University of Missouri-St. Louis. It includes data from the years 2012 to 2022.

The data set from DESE includes the years of 1991 to 2023. In this data set, each public school building in the state of Missouri is represented. In addition, their enrollment numbers for each grade level, pre-school through 12th grade, are listed. There are also summation columns that include enrollment numbers for kindergarten through 8th grade, 9th through 12th grades, kindergarten through 12th grade, and pre-school through 12th grade.

2.1 Data Description

The IPEDS data set is a CSV file that is 1.80 KB in size. This data set has 13 rows and 13 columns. The following attributes in the IPEDS data set were used.

- Institution Name: text; name of the university
- Number of first-time undergraduates - in-state (DRVEF2022): number; first-time freshmen in 2022
- Number of first-time undergraduates - in-state (DRVEF2021): number; first time freshmen in 2021
- Number of first-time undergraduates - in-state (DRVEF2020): number; first-time freshmen in 2020
- Number of first-time undergraduates - in-state (DRVEF2019): number; first-time freshmen in 2019
- Number of first-time undergraduates - in-state (DRVEF2018): number; first-time freshmen in 2018
- Number of first-time undergraduates - in-state (DRVEF2017): number; first-time freshmen in 2017
- Number of first-time undergraduates - in-state (DRVEF2016): number; first-time freshmen in 2016
- Number of first-time undergraduates - in-state (DRVEF2015): number; first-time freshmen in 2015
- Number of first-time undergraduates - in-state (DRVEF2014): number; first-time freshmen in 2014
- Number of first-time undergraduates - in-state (DRVEF2013): number; first-time freshmen in 2013

– Number of first-time undergraduates - in-state (DRVEF2012): number; first-time freshmen in 2012

The DESE data set is an Excel worksheet that is 9.09 MB in size. This data set is much more robust with 98,455 rows and 23 columns. The following attributes from the DESE data set were used.

– EnrollmentGrades09: number; enrollment for 9th grade students
– EnrollmentGrades10: number; enrollment for 10th grade students
– EnrollmentGrades11: number; enrollment for 11th grade students
– EnrollmentGrades12: number; enrollment for 12th grade students

2.2 Data Limitations

The data utilized in this project uses data captured at a census date. The numbers, therefore, may fluctuate due to dis-enrollment. In the elementary and secondary data set, there is no guarantee that a student represented in a particular grade level's enrollment numbers will stay enrolled and graduate at the expected time. Due to this, forecasted enrollment projections may be slightly skewed.

2.3 Data Cleaning and Preparation

The IPEDS data set is clean except for 4 missing values. Harris-Stowe University is missing 2 enrollment values, Missouri State University is missing 1 enrollment value, and Missouri Western State University is missing 1 enrollment value. Additionally, the data set was converted from a CSV file to a XLSX file.

The DESE data set required a more in-depth cleaning for it to be usable. Many of attributes have missing values indicated with an asterisk. These values had to be changed to zeros. There are also many more columns than what are necessary for this project. Grades pre-kindergarten through 12 are included, but only the 9th through the 12th grades are needed. There are also many more years of data included in this data set than the IPEDS data set, and several of the years were removed. Finally, the enrollment numbers for the different grades had to be summed.

2.4 Tools and Techniques for Data Cleaning

Both the DESE and IPEDS data sets did not require extensive cleaning. The DESE data file is an Excel file, and the IPEDS data file was a CSV file. For the purpose of this project, the IPEDS data file was converted into an Excel file. All data cleaning and pre-processing was completed using Excel.

After the file conversion, the IPEDS data set only required the handling of missing values. There were 4 missing values in total. In order to handle those missing values, zeros were inputted into the records.

The DESE data set required more pre-processing and cleaning. The first step was deleting the excess grade levels (PK-8) that were not required for the

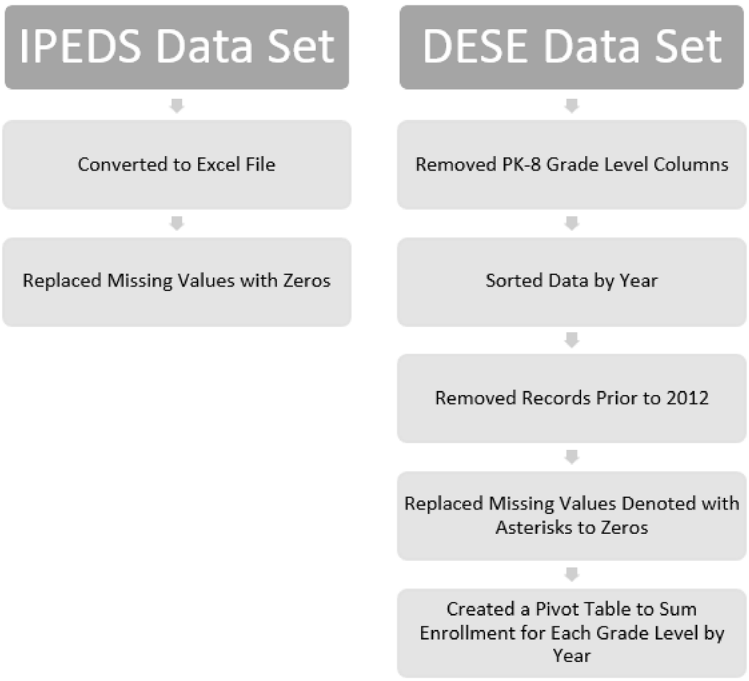

Fig. 1. Data Cleaning Process, Binkley-Hopper 2024

scope of this project. After that, the data was sorted by year, and all data was deleted that were prior to 2012 because the IPEDS data only included the years 2012-2022. Finally, each grade column was sorted, and any asterisks (*) were replaced with zeros. The missing values were handled this way because many of the buildings in the list do not have the 9th-12th grades, so zero enrollment numbers are to be expected.

In addition, the DESE data set was pre-processed with a pivot table in Excel. The project requires total enrollment numbers by grade and by year. A pivot table enabled the data to be quickly gathered. Each column contains a year, and each row represents the sum of enrollment for each grade level.

For the machine learning section of this project, data from the IPEDS data set and DESE data set were combined. For this part of the project, only the data for Northwest Missouri State University was required from the IPEDS data set. This data was combined with the 12th grade enrollment data from the DESE data set in order to build the machine learning models.

3 Exploratory Data Analysis

This project was created after reviewing public data sets on the IPEDS website. Knowing that enrollment is an important metric for higher education institutions, it was determined that the project would focus on enrollment trends. A file was created that included all public, 4-year institutions in the state of Missouri and their first-time freshman enrollment numbers for the last 10 years. This data was analyzed using Microsoft Excel and Python, and it was visualized in Tableau.

3.1 Exploratory Data Analysis with Python

For this step, the IPEDS data set was read into Python utilizing the Pandas library. In addition to the Pandas library, the Matplotlib-Pyplot library was utilized as well. After reading in the IPEDS data set, the first exploratory analysis completed was getting the descriptive statistics of the data set. The descriptive statistics were gathered utilizing the following code in figure 2.

```
ipeds = pd.read_excel("IPEDSData_Transpose.xlsx") #read file
ipeds.set_index('Year', inplace = True) #set column index
ipeds.describe() #Finding descriptive statistics of data set
```

Fig. 2. IPEDS Data Descriptive Statistics Code, Binkley-Hopper 2024
The algorithm in this figure was written to read in an Excel file, set the year as the index, and display descriptive statistics.

The descriptive statistics, table displayed in figure 3, show an interesting snapshot of first-time freshman enrollment statistics at 4-year state universities

in Missouri. For every university represented in this data set, their maximum enrollment happened prior to 2016 except for Missouri Southern State University, Missouri University of Science and Technology, and University of Missouri-Kansas City. Seven of the universities included had their highest enrollment in 2012, which is the first year that is included in this data set. In addition, every university included has experienced their lowest enrollment numbers in the years 2020-2022 except for Missouri University of Science and Technology and University of Missouri-Columbia. This supports the idea that first-time freshman enrollment is on the decline.

	Harris-Stowe State University	Lincoln University	Missouri Southern State University	Missouri State University-Springfield	Missouri University of Science and Technology	Missouri Western State University	Northwest Missouri State University	Southeast Missouri State University	Truman State University	University of Central Missouri	University of Missouri-Columbia	University of Missouri-Kansas City	University of Missouri-St Louis
count	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000
mean	152.181818	294.454545	639.909091	2210.454545	1022.636364	683.363636	839.000000	1275.818182	836.181818	1291.090909	3540.000000	812.727273	370.636364
std	98.420342	74.817596	131.253537	798.592307	136.192711	289.767932	83.190144	142.258088	209.765497	260.511595	411.491677	50.870602	55.968335
min	0.000000	184.000000	423.000000	0.000000	862.000000	0.000000	673.000000	1069.000000	512.000000	904.000000	2739.000000	749.000000	229.000000
25%	92.000000	241.000000	547.500000	2162.500000	911.500000	513.000000	798.500000	1156.000000	653.000000	1046.000000	3237.000000	772.000000	351.500000
50%	168.000000	288.000000	668.000000	2381.000000	1000.000000	760.000000	870.000000	1289.000000	932.000000	1407.000000	3722.000000	805.000000	378.000000
75%	227.000000	343.500000	737.500000	2705.000000	1148.000000	887.000000	893.500000	1396.000000	993.000000	1482.500000	3792.000000	837.000000	405.500000
max	277.000000	438.000000	799.000000	2811.000000	1211.000000	990.000000	926.000000	1485.000000	1043.000000	1597.000000	4096.000000	907.000000	442.000000

Fig. 3. IPEDS Data Descriptive Statistics, Binkley-Hopper 2024
This table shows descriptive statistics for public 4-year universities in the state of Missouri. These statistics include the mean enrollment, the standard deviation, the minimum enrollment, and the maximum enrollment for each university.

The full exploratory data analysis completed in Python can be viewed online in this project's Github Repo.

3.2 Exploratory Data Analysis with Excel

The IPEDS data was further analyzed creating a visualization in Excel. In figure 4, Enrollment Numbers of First-Time Freshmen, Each university's enrollment numbers for each year from 2012 to 2022 is displayed in a line chart. By viewing this visualization, it is verified that enrollment for every university has declined since 2012. This further supports the conclusion that enrollments at universities are declining.

The full Excel file with data and visualization included can be downloaded from this project's Github Repo.

3.3 Exploratory Data Analysis Visualizations in Tableau

Several visualizations were created in Tableau in order to analyze the data further. The first visualization that was created was a treemap. The treemap, in figure 5, shows total summed enrollment at public 4-year universities in the state of Missouri by year. Each year is represented on the treemap with a box. The boxes are coded by size and color. The larger green boxes have the higher enrollments. The smaller yellow, orange, and red boxes have the lower enrollments. This treemap shows the trend of enrollment declining from 2022-2012.

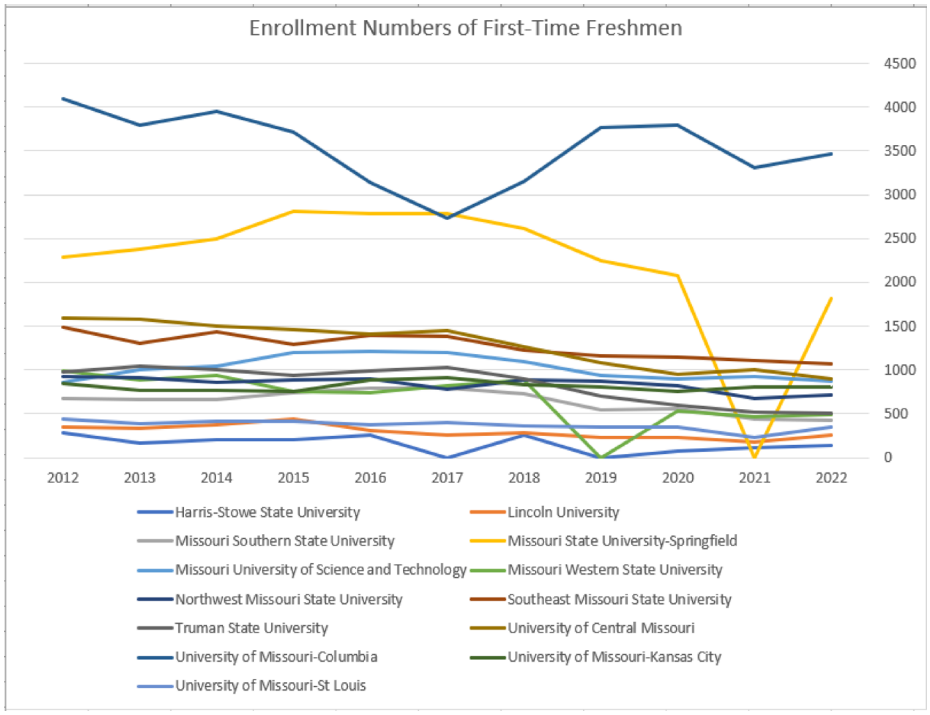

Fig. 4. First-Time Freshman Enrollment, Binkley-Hopper 2024
This chart shows the enrollment trends of first-time freshmen from the state of Missouri at public 4-year universities in Missouri. This chart verifies the assertion that enrollment of first-time freshmen has declined in the state of Missouri since 2012.

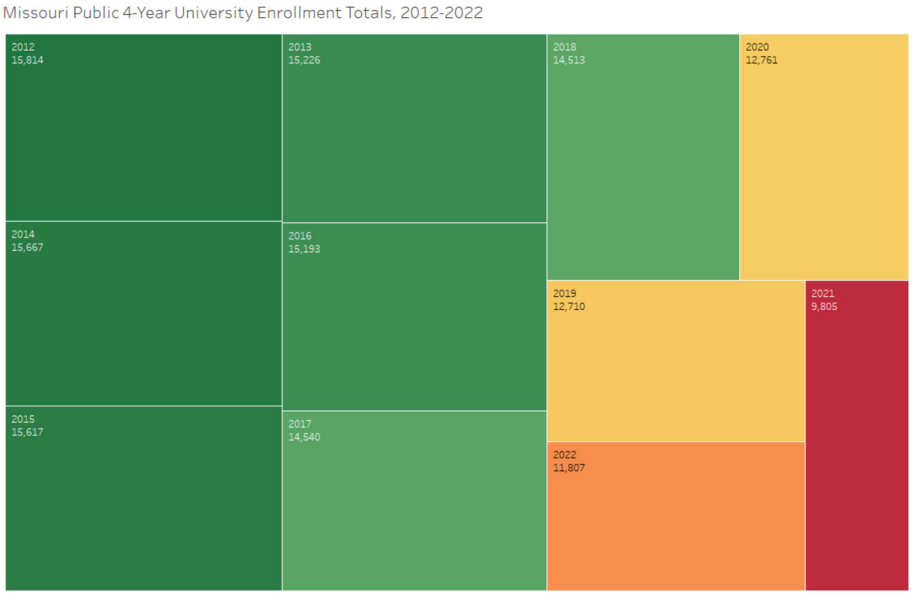

Fig. 5. Missouri Public 4-Year University Enrollment Totals, 2012-2022, Binkley-Hopper 2024

This treemap depicts the sum of university enrollment at public 4-year institutions in the state of Missouri. The boxes in green have the highest enrollment numbers, and the boxes in yellow, orange, and red have the lowest enrollment numbers. It shows the overall decline of university students at the public 4-year universities from the years 2012-2022.

The next set of visualizations created in Tableau, shown in figures 6 and 7, feature each Missouri public 4-year university's enrollment from 2012-2022 compared to 12th grade enrollment during the same years as reported by DESE. These visualizations all show a trend of declining freshman enrollment from 2012-2022.

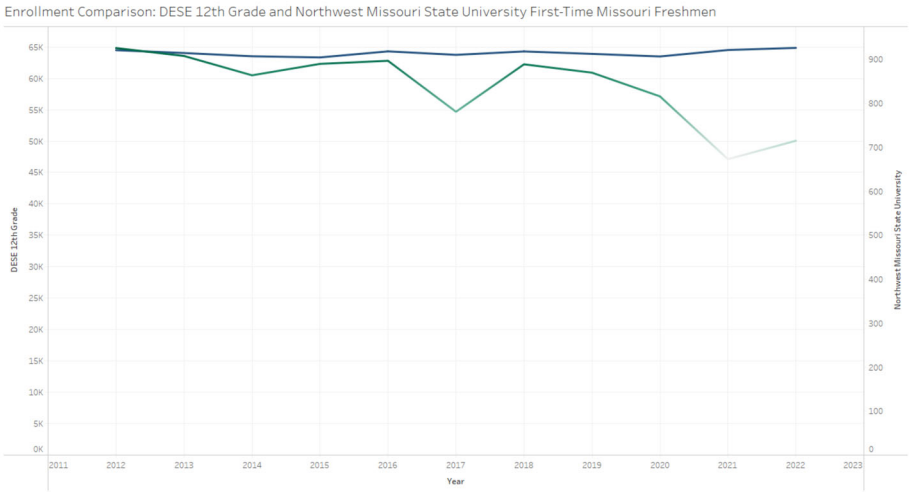

Fig. 6. Enrollment Comparison: DESE 12th Grade and Northwest Missouri State University First-Time Missouri Freshmen, Binkley-Hopper 2024
This line chart compares the enrollment trends of Missouri 12th grade students as well as first-time Missouri freshmen at Northwest Missouri State University. This chart shows that enrollment at Northwest has declined in recent years.

The full exploratory data analysis completed in Tableau can be viewed online in this project's Tableau Public Site

4 Model Building

In addition to the exploratory data analysis completed in section 3, machine learning forecasting algorithms were applied to the data set for Northwest Missouri State University. In order to complete the algorithms, data to train the model was required. A data set was sourced from the DESE website that includes enrollment numbers from every public-school building in the state of Missouri by grade. This data was processed utilizing Python to build the machine learning algorithms.

This model uses machine learning to analyze the following problem. How do high school enrollment numbers in Missouri public high schools impact enrollment numbers of first-time freshmen at Northwest Missouri State University?

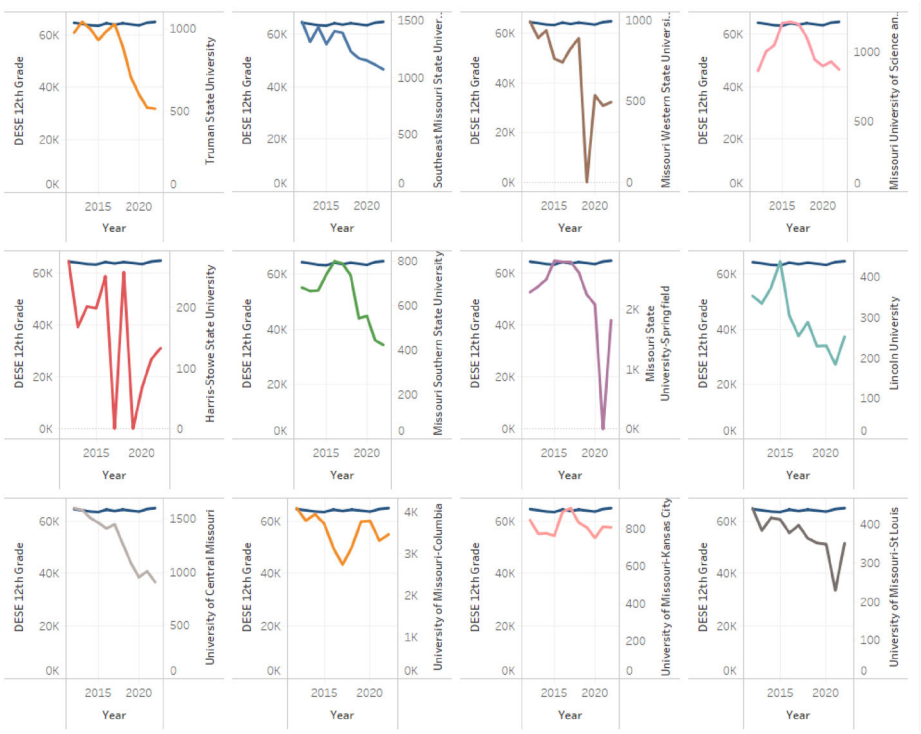

Fig. 7. Enrollment Comparison: DESE 12th Grade and Missouri Public 4-year University First-Time Missouri Freshmen, Binkley-Hopper 2024
These line charts compare the enrollment trends of Missouri 12th grade students as well as first-time Missouri freshmen at Harris-Stowe State University, Lincoln University, Missouri Southern State University, Missouri State University, Missouri University of Science and Technology, Missouri Western State University, Southeast Missouri State University, Truman State University, University of Central Missouri, University of Missouri-Columbia, University of Missouri-Kansas City, and University of Missouri-St. Louis. These charts show that enrollment has declined in recent years.

To answer this problem, forecasting was utilized to attempt to provided predictions on first-time freshmen enrollment numbers for students from the state of Missouri. The independent and dependent variables are as follows.

- **Independent Variable:** high school enrollment numbers in Missouri public high schools.
- **Dependent Variable:** Missouri first-time freshman enrollment numbers at Northwest Missouri State University.

The machine learning models that were used for analyzing the data to solve the problem were a linear regression model, a polynomial regression model, and an elastic net with polynomial features model. The data set in figure 8 was read into each model using Python. The training/testing split was accomplished using the train/test/split module from the Scikit-Learn library in Python (see figure 9). The amount of data used for the training set was 80 percent, and the amount of data used for the testing set was 20 percent.

Year	NW_Enroll	DESE12_Enroll
2012	926	64491
2013	908	64043
2014	864	63528
2015	890	63343
2016	897	64306
2017	781	63752
2018	889	64294
2019	870	63907
2020	816	63496
2021	673	64524
2022	715	64862

Fig. 8. Machine Learning Data Set, Binkley-Hopper 2024
Data was compiled from the years 2012-2022 from DESE and IPEDS. The DESE data is total 12th grade enrollment numbers for the state of Missouri. The IPEDS data is Missouri first-time freshman enrollment totals for Northwest Missouri State University.

The full machine learning modeling completed in Python can be viewed online in this project's Github Repo.

4.1 Linear Regression Model

The first machine learning model created for this project was a linear regression model. Linear regression is a machine learning model that seeks to find the

```
from sklearn.model_selection import train_test_split

train_set, test_set = train_test_split(nwDESE,
                                       test_size=0.2, random_state=123)
print('Train size: ', len(train_set), 'Test size: ', len(test_set))
```

[3]

```
Train size:  8 Test size:  3
```

Fig. 9. Train/Test Split Code, Binkley-Hopper 2024
The train/test/split module from the Scikit-Learn library in Python was used to split the data into training and testing sets. The standard 80/20 split was utilized in this project. 80 percent of the data was used for the training of the model, and 20 percent of the data was used for the testing of the model.

value of a dependent variable (y) based on the independent variable (x). This is accomplished by fitting a straight line to the model in the hope of predicting y from x.[3]

In order to complete the linear regression model, the following libraries and modules were used in Python.

- Matplotlib: Pyplot
- Numpy
- Pandas
- Scikit-Learn: Linear Model (Linear Regression)
- Scikit-Learn: Metrics (Mean Absolute Error, Mean Squared Error, and R2 Score)

Training/Testing the Linear Regression Model The training and testing of the linear regression model was started by importing necessary libraries. Variables were then assigned to have x and y values that were used both in training and testing. After this, the linear regression model was implemented. This process can be viewed in figure 10.

The results of the linear regression model can be viewed in figure 11. In the training data, the Root Mean Square Error was 62.82, and the R-squared value was .41. In a perfect model, the Root Mean Square Error would be 0, and the R-squared value would be 1. The values calculated show that the results of the linear regression on the training data was not significant. In the testing data, the Root Mean Squared Error was 112.47, and the R-squared value was -2.22. The results of the linear regression on the testing data were also not significant.

Plotting the Linear Regression Model An algorithm was also written to plot the results of the linear regression on the data set. After importing Pyplot

```
X = train_set[['DESE12_Enroll']]
y = train_set['NW_Enroll']
X_test = test_set[['DESE12_Enroll']]
y_test = test_set['NW_Enroll']

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))
```

Fig. 10. Linear Regression Code, Binkley-Hopper 2024
This figure shows the algorithm used to perform a linear regression on the data set in figure 8. The dependent variable (y) is Northwest first-time freshman enrollment from the state of Missouri, and the independent variable (x) is the 12th grade enrollment in Missouri as reported by DESE. These variables were used in both training and testing sets, and a linear regression model was created. The last part of the code was written to display the results of the machine learning algorithm performed.

```
...   Results for linear regression on training data
       Default settings
      Internal parameters:
       Bias is  7561.386276211434
       Coefficients [-0.10520782]
       Score 0.4149514449171704
      MAE is  52.10351695793656
      RMSE is  62.82008242479604
      MSE is  3946.362755858169
      R^2  0.4149514449171704

      Results for linear regression on test data
      MAE is  107.95220963401214
      RMSE is  112.47126337128135
      MSE is  12649.785084332136
      R^2  -2.22314890886669
```

Fig. 11. Linear Regression Output, Binkley-Hopper 2024
The results for both the training and the testing data that were used in the machine learning linear regression model were both insignificant. The RMSE values are too high and the R-Squared values are too low.

from Matplotlib, the code was written to plot the results. This code can be viewed in figure 12.

The linear regression plot created confirms the results of the linear regression. As can be seen in figure 13, the linear regression line is not a good fit for the plotted data points. The regression line does not intersect any of the data points, and most of the data points are not close to the regression line.

4.2 Polynomial Regression Model

The next regression model that was created for this project was a polynomial regression model. Like the linear regression model, the polynomial regression model looks to find the value of a dependent variable based on the independent variable. The difference with this model is that is uses a polynomial, or curved, line to fit the data.[6]

In order to complete the polynomial regression model, the following libraries and modules were used in Python.

- Matplotlib: Pyplot
- Numpy
- Pandas
- Scikit-Learn: Linear Model (Linear Regression)

```
min_dese12Enroll = dese12Enroll.min()
max_dese12Enroll = dese12Enroll.max()
points = 200
step_by = (max_dese12Enroll - min_dese12Enroll)/(points-1)

x_values = [min_dese12Enroll + i*step_by for i in range(0, points)]
inputs = [[x] for x in x_values]
y_values = lr_model.predict(inputs)

plt.scatter(dese12Enroll, nwEnroll, c=year)
plt.xlabel('DESE 12th Grade Enrollment')
plt.ylabel('Northwest Freshman Enrollment')
plt.plot(x_values, y_values, c='red')
plt.legend()
plt.show()
```

Fig. 12. Linear Regression Plot Code, Binkley-Hopper 2024
This algorithm was created to plot the results from the linear regression that was performed. It was the goal to create a scatter plot that graphed the DESE totals (x) and Northwest totals (y).

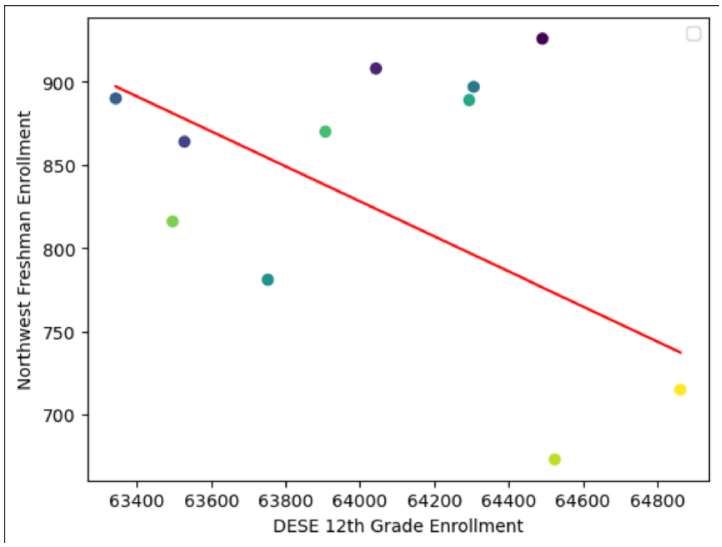

Fig. 13. Linear Regression Plot, Binkley-Hopper 2024
This plot represents the linear regression that was performed on the DESE 12th grade enrollment and Northwest freshman enrollment data set. The plot confirms that the results of the linear regression analysis were insignificant. As can be seen in this plot, none of the data points are intersected by the linear regression line, and most of the data points are not close to it either.

- Scikit-Learn: Metrics (Mean Absolute Error, Mean Squared Error, and R2 Score)
- Scikit-Learn: Preprocessing (PolynomialFeatures)

Training/Testing the Polynomial Regression Model Like the linear regression model, the training and testing of the polynomial regression model was started by importing necessary libraries. The degree of the polynomial was set to 8 in this model. Variables were then assigned to have x and y values that were used both in training and testing. In addition, a polynomial variable was created for each the training and testing phases. With the polynomial variables established, the polynomial regression was able to be implemented. The polynomial regression model algorithm can be viewed in figure 14.

The output of the polynomial regression model can be viewed in figure 15. The polynomial regression model provided the best fit out of the 3 machine learning models used for this project. However, the results still show that this model was insignificant as well. The Root Mean Squared Error was 258.61 for the training set and 478.77 for the test set, which are very high. The R-squared value for the training set was .74 and -1.33 for the test set. While the R-squared value for the training set was better than any of the other R-squared values, it was still insignificant and all of the other values were insignificant as well.

Plotting the Polynomial Regression Model The algorithm produced to plot the polynomial regression model used the same x and y values as the linear regression model. The code deviated from the linear regression model by incorporating polynomial features. This algorithm can be viewed in figure 16.

The polynomial regression plot, shown in figure 17, again confirms the earlier analysis that the results are insignificant. The polynomial line is a better fit for the data, but there are still many outliers. Therefore, the polynomial regression model can not be used to make an accurate prediction of future Northwest freshman enrollment.

4.3 Elastic Net with Polynomial Features Model

The final machine learning model used in this project was an elastic net with polynomial features model. An elastic net model combines features from both the linear regression model and polynomial model to find the value of the dependent variable based on the independent variable.

In order to complete the elastic net with polynomial features model, the following libraries and modules were used in Python.

- Matplotlib: Pyplot
- Numpy
- Pandas
- Scikit-Learn: Linear Model (Elastic Net and Linear Regression)
- Scikit-Learn: Metrics (Mean Absolute Error, Mean Squared Error, and R2 Score)
- Scikit-Learn: Preprocessing (PolynomialFeatures)

```
X = train_set[['NW_Enroll']]
y = train_set['DESE12_Enroll']
X_poly = poly_process.fit_transform(X)

X_test = test_set[['NW_Enroll']]
y_test = test_set['DESE12_Enroll']
X_poly_test = poly_process.fit_transform(X_test)

lr_model = LinearRegression()
lr_model.fit(X_poly,y)

y_pred = lr_model.predict(X_poly)
print('Results for linear regression on training data')
print('Polynomial regression with degree ', power)
print(' Default settings')
print('Internal parameters:')
print(' Bias is ', lr_model.intercept_)
print(' Coefficients', lr_model.coef_)
print(' Score', lr_model.score(X_poly,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_poly_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))
```

Fig. 14. Polynomial Regression Code, Binkley-Hopper 2024
This figure shows the algorithm used to perform a polynomial regression on the data set in figure 8. The degree of this polynomial regression was 8. The dependent variable (y) is Northwest first-time freshman enrollment from the state of Missouri, and the independent variable (x) is the 12th grade enrollment in Missouri as reported by DESE. In addition polynomial variables were created for both the training and testing sets. These variables were used in both training and testing sets, and a polynomial regression model was created. The last part of the code was written to display the results of the machine learning algorithm performed.

```
Results for linear regression on training data
Polynomial regression with degree  8
 Default settings
Internal parameters:
 Bias is  12954.895516611636
 Coefficients [ 1.66593908e-19 -5.47560334e-13  8.86201316e-14  3.73503659e-11
  9.24721202e-09 -2.86626210e-11  3.02903285e-14 -1.08363000e-17]
 Score 0.7355353297931051
MAE is  172.48568705757498
RMSE is  258.60503758139674
MSE is  66876.56546247563
R^2  0.7355353297931051

Results for linear regression on test data
MAE is  423.1622734557216
RMSE is  478.7712482866863
MSE is  229221.9081859918
R^2  -1.325127384779159
```

Fig. 15. Polynomial Regression Output, Binkley-Hopper 2024
The results for both the training and the testing data that were used in the machine learning polynomial regression model were both insignificant. The RMSE values are too high and the R-Squared values are too low.

Training/Testing the Elastic Net with Polynomial Features Model The elastic net with polynomial features model, like the others, started with importing the necessary libraries. The variables for this model included training and testing sets of x and y variables, as well as polynomial x values for both the training and testing sets. These variables were inputted into an elastic net model, and the results printed. This model is shown in figure 18.

The results of the elastic net model with polynomial features can be viewed in figure 19. The training set's Root Square Mean Error was 58.45, and the R-squared value was .49. The test set had a Root Square Mean Error of 110.26 and a R-squared value of -2.10. These values were all insignificant, and, therefore, this model was not a good fit.

Plotting the Elastic Net Polynomial Features Model The algorithm produced to plot the elastic net model with polynomial features used the same x and y values as the other models. The code was similar to the code used in the polynomial regression plot. This code sample is shown in figure 20.

The plot produced by the elastic net with polynomial features, shown in figure 21, confirms that the elastic net model is not a good fit for predicting freshman enrollment at Northwest. The majority of the data points on this plot do not fit the regression line.

```
nwEnroll = nwDESE['NW_Enroll']
deseEnroll = nwDESE['DESE12_Enroll']
year = nwDESE['Year']

min_enroll = nwEnroll.min()
max_enroll = nwEnroll.max()
points = 200
step_by = (max_enroll - min_enroll)/(points-1)

x_values = [min_enroll + i*step_by for i in range(0, points)]
inputs = [[x] for x in x_values]
inputs_poly = poly_process.fit_transform(inputs)
y_values = lr_model.predict(inputs_poly)

plt.scatter(nwEnroll, deseEnroll, c=year)
plt.xlabel('Northwest Freshman Enrollment')
plt.ylabel('DESE 12th Grade Enrollment')
plt.plot(x_values, y_values, c='red')
plt.show()
```

Fig. 16. Polynomial Regression Plot Code, Binkley-Hopper 2024
This algorithm was created to plot the results from the polynomial regression that was performed. It was the goal to create a scatter plot that graphed the DESE totals (x) and Northwest totals (y). This code also accounts for the polynomial line used to fit the data on the plot.

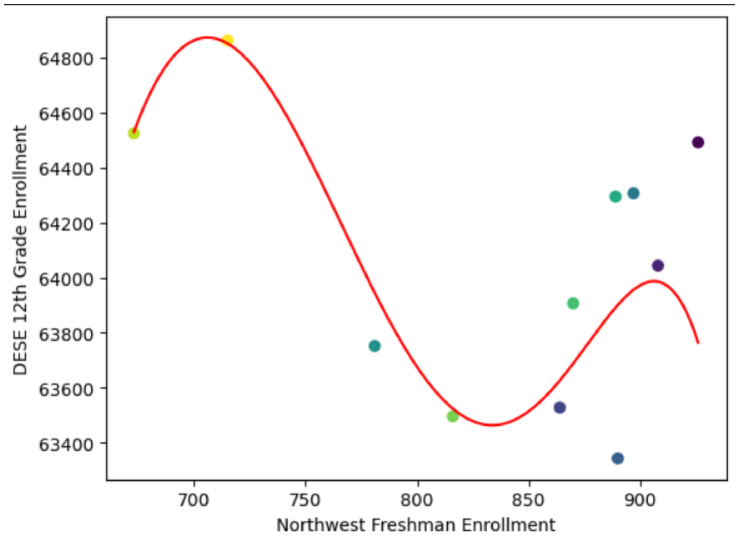

Fig. 17. Polynomial Regression Plot, Binkley-Hopper 2024
This plot represents the polynomial regression that was performed on the DESE 12th grade enrollment and Northwest freshman enrollment data set. The plot confirms that the results of the polynomial regression analysis were insignificant. As can be seen in this plot, while a better fit than the linear regression line, there are still many outliers with the polynomial regression line.

```
X = train_set[['DESE12_Enroll']]
y = train_set['NW_Enroll']
X_poly = poly_process.fit_transform(X)

X_test = test_set[['DESE12_Enroll']]
y_test = test_set['NW_Enroll']
X_poly_test = poly_process.fit_transform(X_test)

reg_lr_model = ElasticNet(alpha=0.3, l1_ratio=0.5)
reg_lr_model.fit(X_poly,y)

y_pred = reg_lr_model.predict(X_poly)
print('Results for elastic net on training data')
print('Polynomial regression with degree ', power)
print(' Default settings')
print('Internal parameters:')
print(' Bias is ', reg_lr_model.intercept_)
print(' Coefficients', reg_lr_model.coef_)
print(' Score', reg_lr_model.score(X_poly,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 = reg_lr_model.predict(X_poly_test)
print()
print('Results for elastic net 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))
```

Fig. 18. Elastic Net Code, Binkley-Hopper 2024
This figure shows the algorithm used to perform an elastic net with polynomial features model on the data set in figure 8. Like the polynomial regression, this model used a degree of 8. The dependent variable (y) is Northwest first-time freshman enrollment from the state of Missouri, and the independent variable (x) is the 12th grade enrollment in Missouri as reported by DESE. In addition, polynomial variables were created for both the training and testing sets. These variables were used in both training and testing sets, and an elastic net with polynomial features model was created. The last part of the code was written to display the results of the machine learning algorithm performed.

```
Results for elastic net on training data
Polynomial regression with degree  8
 Default settings
Internal parameters:
 Bias is  -57522.860785034354
 Coefficients [ 1.24122554e+00 -1.49697525e-06 -1.56468519e-11 -1.83132489e-16
 -2.28598489e-21 -2.97201346e-26 -3.97377457e-31 -5.42314235e-36]
 Score 0.4934741509045908
MAE is  45.70591806033099
RMSE is  58.45253531606267
MSE is  3416.6988848755536
R^2  0.4934741509045908

Results for elastic net on test data
MAE is  106.41575261166629
RMSE is  110.25679076565018
MSE is  12156.559909940364
R^2  -2.0974757711755645
```

Fig. 19. Elastic Net Output, Binkley-Hopper 2024
The results for both the training and the testing data that were used in the machine learning elastic net with polynomial features model were both insignificant. The RMSE values are too high and the R-Squared values are too low.

```
dese = nwDESE['DESE12_Enroll']
nw = nwDESE['NW_Enroll']
year = nwDESE['Year']

min_dese = dese.min()
max_dese = dese.max()
points = 200
step_by = (max_dese - min_dese)/(points-1)

x_values = [min_dese + i*step_by for i in range(0, points)]
inputs = [[x] for x in x_values]
inputs_poly = poly_process.fit_transform(inputs)
y_values = reg_lr_model.predict(inputs_poly)

plt.scatter(dese, nw, c=year)
plt.xlabel('DESE 12th Grade Enrollment')
plt.ylabel('NW Freshman Enrollment')
plt.plot(x_values, y_values, c='red')
plt.show()
```

Fig. 20. Elastic Net Plot Code, Binkley-Hopper 2024
This algorithm was created to plot the results from the elastic net model with polynomial features that was performed. It was the goal to create a scatter plot that graphed the DESE totals (x) and Northwest totals (y). This code also accounts for the elastic net with polynomial features line used to fit the data on the plot.

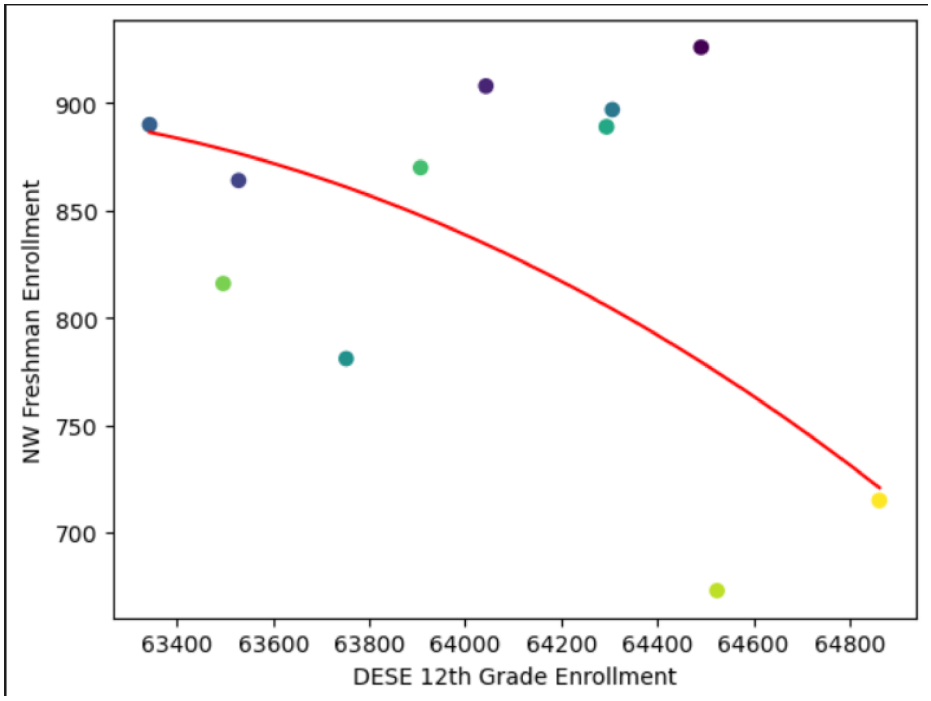

Fig. 21. Elastic Net Plot, Binkley-Hopper 2024
This plot represents the elastic net with polynomial features model that was performed on the DESE 12th grade enrollment and Northwest freshman enrollment data set. The plot confirms that the results of this analysis were insignificant. As can be seen in this plot, the regression line is not a good fit for the data points. Almost all of the data points on this plot are outliers with the regression line only fitting 2 values.

5 Conclusion

This project was started with two goals in mind. The first goal was to examine and compare enrollment trends for public 4-year institutions in the state of Missouri. This was accomplished through completing an exploratory data analysis with Python, Excel, and Tableau. The second goal was to examine the enrollment trends for Northwest Missouri State University and use them to train a machine learning model in Python to forecast future enrollment projections. Linear regression, polynomial regression, and elastic net model with polynomial features were the machine learning models used for the second goal.

The exploratory data analysis performed for this project sought to examine and compare enrollment trends for public 4-year institutions in the state of Missouri. This project accomplished this goal. The first type of exploratory analysis completed was finding the descriptive statistics of the data set with Python. The descriptive statistics confirmed the assertion that first-time freshman enrollment is on the decline. The next type of exploratory analysis completed was creating a line chart to show the enrollment numbers of the same universities from the years 2012-2022. This chart also showed a downward trend for freshman enrollment. Finally, Tableau was used to create a tree map and individual line charts. The tree map shows the decline of total Missouri public 4-year university enrollment, and the line charts show the comparison of each of the universities represented in the data set with 12th grade enrollment as reported by DESE.

The exploratory data analysis provided an overall picture of the data sets that were being used. The enrollment trends for each of the included universities showed how the enrollment of freshmen in the state of Missouri was on the decline, and each of the analyses reinforced the conclusions drawn. The exploratory data analysis provided the foundation for the machine learning model building that was required for the second goal.

Three machine learning models, linear regression, polynomial regression, and elastic net with polynomial features, were used to try to find a trend in enrollment at Northwest Missouri State University so that future enrollment could be predicted. The results of these models, shown in table 1, did not provide any useful correlation.

Table 1. Machine Learning Model Results Summary, M. Binkley-Hopper 2024

Model	Training RMSE	Training R-Squared	Testing RMSE	Testing R-Squared
Linear Regression	62.82	0.41	112.47	-2.22
Polynomial Regression	258.61	0.74	478.77	-1.33
Elastic Net	58.45	0.49	110.26	-2.10

There are a few different possibilities that could explain the failure of the machine learning models utilized in this project. The first possible explanation is that the sample size was too small. This project only utilized data from 2012-2022. These years are what is publicly available through IPEDS.

Another possible explanation for the failure of the models, is the Covid-19 pandemic from 2020-2021. According to Fields and Brint, there was a disruption in enrollment in the United States during the height of the pandemic. [2] This disruption can also be seen in Northwest's enrollment data.

Finally, in recent years, Northwest has started many online degree programs in an effort to combat shrinking enrollment numbers. This initiative could also skew any potential enrollment predictions from machine learning models.

It is the recommendation that any future research takes the following into consideration. There should be a larger sample size utilized. A larger data sample would allow for more training and testing instances in the models. In addition, researchers should decide on a way to handle any outliers, such as the 2020-2021 academic year due to the Covid-19 pandemic.

In conclusion, data and research shows that enrollment is on the decline, both in Missouri and the United States. It is important for higher education institutions to realize this trend and to work to find ways to make up for the enrollment deficits. In addition, it would be advantageous for higher education institutions to consider future research into creating a machine learning model to make enrollment projections.

References

1. CDC: Births., https://www.cdc.gov/nchs/hus/topics/births.htm
2. Fields, B., Brint, S.: The disruption in u.s. public higher education enrollments, 2009–2019: Sources of inter-state variation by tier. Journal of Higher Education **94**(2), 256 – 285 (2023), https://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=afh&AN=162237918&site=eds-live&scope=site&custid=074-800
3. IBM: What is linear regression?, https://www.ibm.com/topics/linear-regression#:~:text=the%20next%20step-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable.
4. Johnstone, S.M., Schexnider, A.J.: Higher education's actual responses to shifting demographics, so far. Change **55**(1), 5 – 16 (2023), https://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=eue&AN=161688034&site=eds-live&scope=site&custid=074-800
5. Knox, L.: Growing enrollment, shrinking future. (2023), https://www.insidehighered.com/news/admissions/traditional-age/2023/10/26/undergraduate-enrollment-first-time-2020
6. Pant, A.: Introduction to linear regression and polynomial regression. (2019), https://towardsdatascience.com/introduction-to-linear-regression-and-polynomial-regression-f8adc96f31cb
7. Trellis, C., Schuette, A.: Navigating the enrollment cliff in higher education. spotlight report brief. (2023), https://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=eric&AN=ED628984&site=eds-live&scope=site&custid=074-800