College Major Popularity Prediction: Survey of Machine Learning Algorithms

Erica Chen

1. Introduction

Prediction of college major popularity rates is an interesting and important problem in the field of education. Accurately predicting the popularity of majors can help universities and colleges plan their curriculum and resources more efficiently. Machine learning algorithms can be used to predict the popularity based on various factors such as demographics, family income-level, etc.

In this project I collected college major data on over 1 million undergraduates over a period of time. I used this data with the intention to explore the relationship between time and college major popularity. The question I wanted to answer was how can historical data on college majors be predicted. I believe college majors are an indicator of success for the long term economy, as well as an artifact to historic events that could affect participation in certain majors such as computer science versus home-economics.

In this survey, we will explore supervised learning technique such as linear regression and decision tree, and evaluate their performance on a data set of total number of college majors over time. I will run a time series analysis on the raw data of 33 majors over 50 years. Then, I will pre process the data to ensure that it is suitable for machine learning algorithms, and then train and test the algorithms on the data. I will compare the Mean Absolute Error (MAE) and Mean Squared Error (MSE) of these algorithms and measure how well the model fits the training data.

This survey aims to provide insights into the effectiveness of various machine learning algorithms for predicting college major popularity and identify the most accurate and efficient algorithm for this task. The results of this survey can be useful for universities and colleges to plan their curriculum and resources more effectively to anticipate the future.

2. Data

2.1. **Data Collection.** The data collection process involves identifying the topic of interest: education over time. The data source for this analysis was retrieved from the National Center for Education Statistics. Many of the data sources from this site focus on variables such as race, gender, year, major, degree type and the like. The site is easy to access and data is exported in excel (XLS) format.

2.2. **Data Exploration.** The raw data shows us that within the cluster of 33 college majors, we see a handful that show consistent increase in popularity such as the Business major. To explore this data set further, I will eliminate noise from least consistent college majors and prioritize based on the majors that show greatest variation through time.

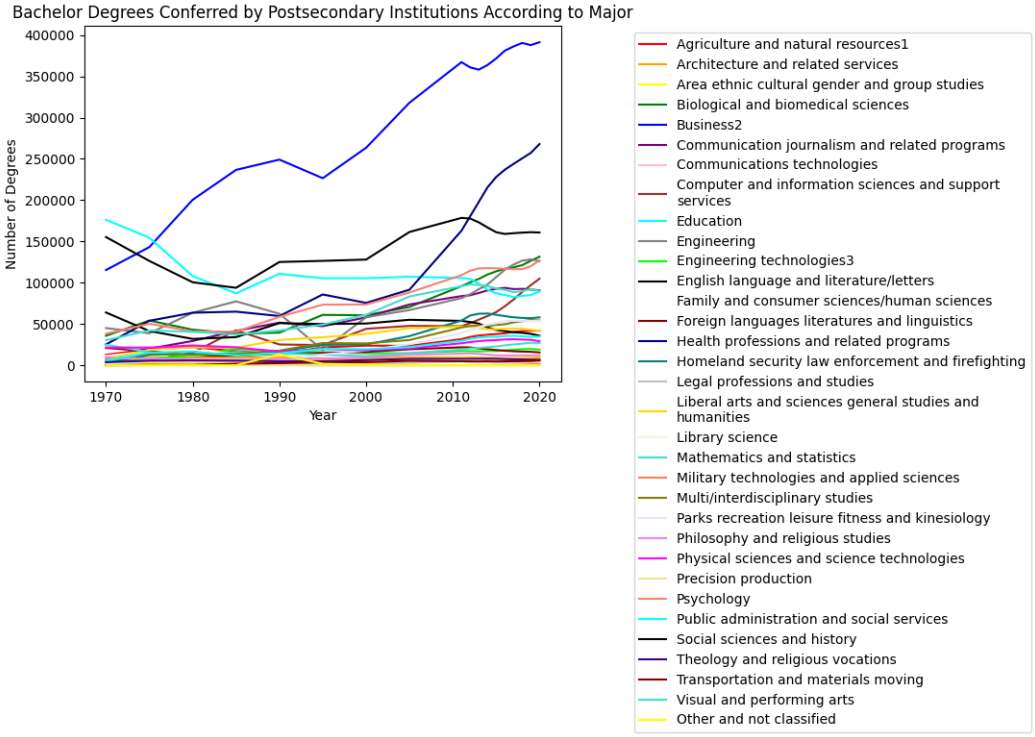

Figure 1. Time series on participation of 33 college majors

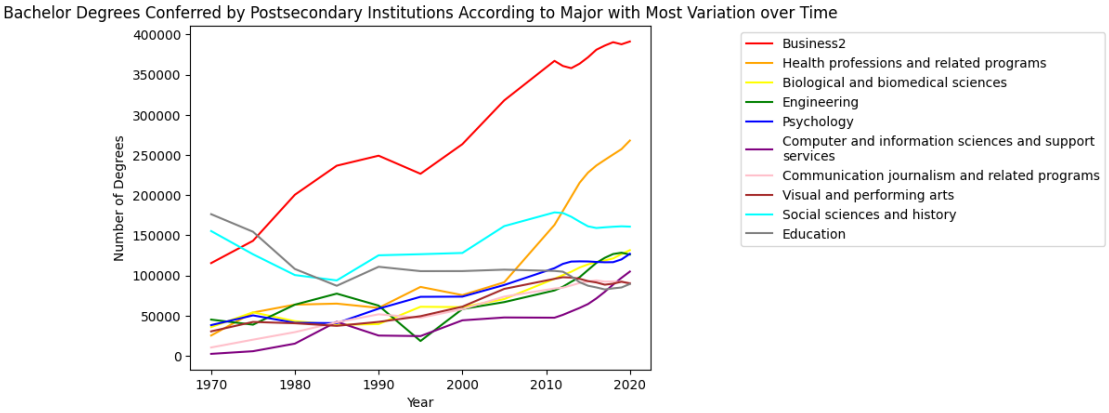

Figure 2. Time Series on top 10 majors with most variation (growth)

2.3. **Data Cleaning.** Reformatting data is an important step in preparing data because it ensures that the data is consistent, accurate, and in the correct format, which allows the machine learning algorithms to produce accurate results.

For linear regression and decision tree algorithms. These algorithms require data to be in a certain format to work effectively. For linear regression, the data should be continuous, and the target variable should be numerical, which I have established to be the count of the total number of degrees conferred by institutions each year. The independent variables should be numeric, and any categorical variables need to be converted into numerical values using dummy encoding or other techniques.

For decision tree algorithms, the data should be split into a series of binary decisions based on the independent variables. This requires converting categorical variables into binary values and transforming continuous variables into a series of binary decisions. Additionally, decision trees require a certain level of data balance to prevent over fitting, so data may need to be under sampled or over sampled to achieve balance.

2.4. **Data Description.** The variables of interest was finalized after much preliminary data analysis. The raw data I am working with encompassed information on the number of Bachelors degrees conferred by post secondary institutions based on field of study (major) over a 50 year period for all U.S. based institutions.

3. Statistical Models

3.1. **Linear Regression.** Linear regression is a popular choice for modeling relationships between continuous variables. It assumes that the relationship between the dependent variable and the independent variables is linear, and tries to find the best-fit line that minimizes the sum of squared errors between the predicted and actual values.

Linear regression is computationally efficient and easy to interpret, making it a good choice for simple problems with a small number of predictors. It is for this reason that I chose to run a linear regression because this data set has fewer variables. The axis in the figures represent continuous variables, where total count of degree represents the y labels and years represents the independent variable. The blue dots are raw data points from our data set. The red line represents our linear regression. Intuitively, the count of high education degrees is expected to increase over time.

It is also important to acknowledge there are intervals of time where data is systematically left out due to lack of information and consistent record keeping, which can result in slight bias. Further exploration of a more specialized imputation technique can mitigate the absence of these patches of data.

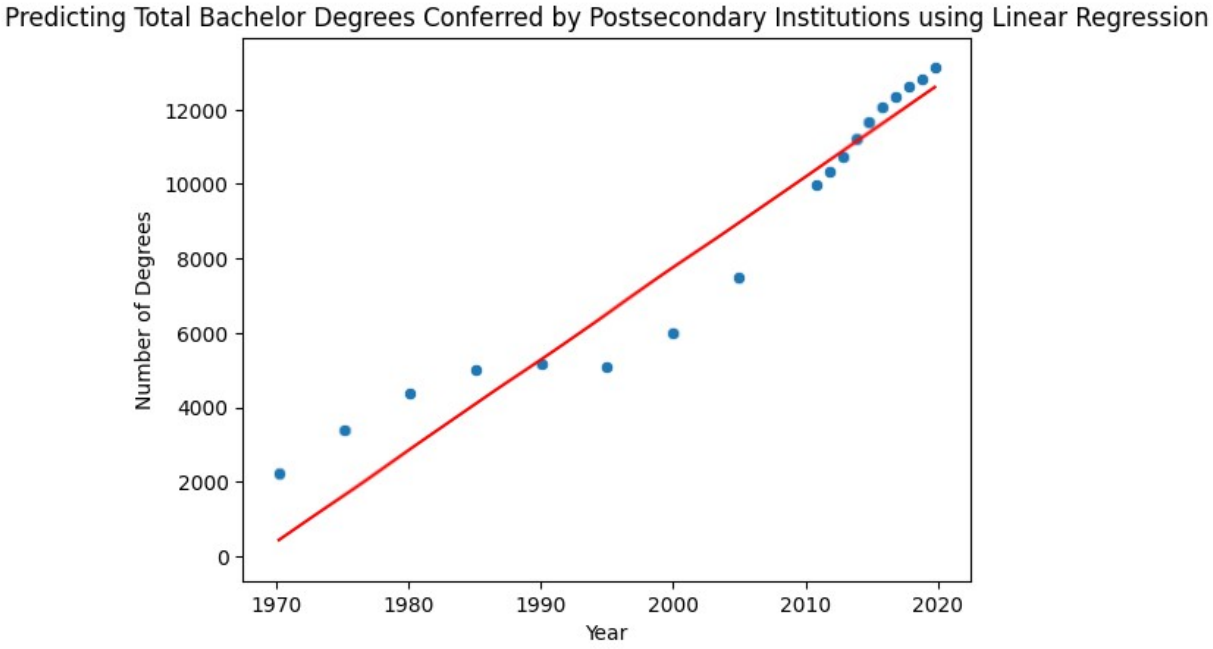

Figure 3. Linear regression on total count of Bachelors degree over time

$$\hat{y} = -34365996.87844855 + 17510.32944203554 * X \tag{3.1}$$

3.2. **Decision Tree.** Decision tree models are a popular choice for classification and regression problems. Decision trees recursively split the data based on the features that are most informative for the target (y label) variable. However, decision trees can be prone to over fitting, especially when the tree is deep or the data set is noisy. To avoid over fitting, pruning techniques such as post-pruning or pre-pruning can be used.

To represent the results of the decision tree, I plot a confusion matrix to show the number of true positive, true negative, false positive, and false negative predictions made by a classification model. Two dark squares in the upper left and lower right corners of the matrix represent the correct classifications, while the two dark squares in the upper right and lower left corners represent the misclassifications. The figure above shows four dark squares along the left diagonal, which suggests that the model is having difficulty distinguishing between the two classes and is making incorrect predictions.

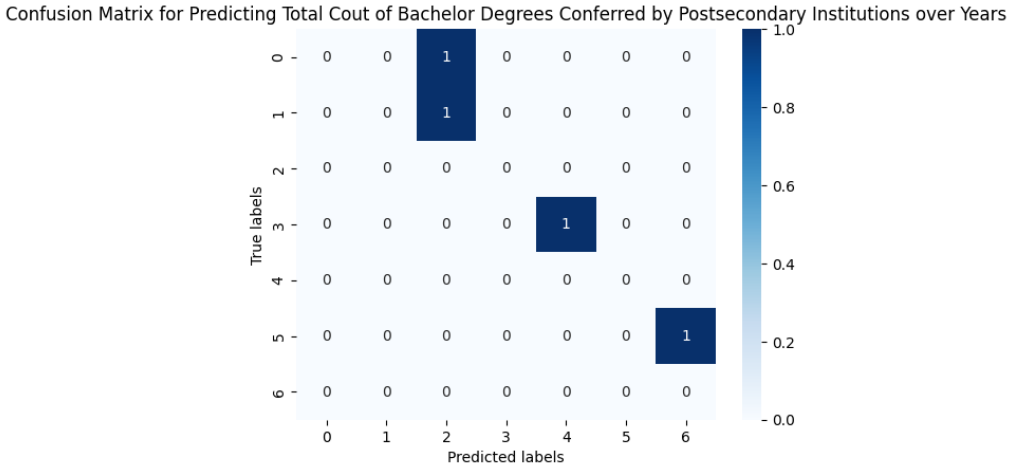

Figure 4. Confusion matrix on the relationship between Year and Total Number of Degrees

4. Results

4.1. **Performance of Models.** I chose to use Mean Absolute Error (MAE) and Mean Squared Error (MSE) to as a metric for success in my survey of machine learning algorithms because they are good metrics for the quality of regression models, which predict continuous numerical values rather than discrete categories. While it is important to acknowledge that MSE is more sensitive to large errors because it squares the differences, MAE treats all errors equally, so the two are a balanced pair of metrics.

MAE values for three models

Model 1: Linear Regression	Model 2: Decision Tree
80782.29254618038	63491.25

Figure 5. Table of MAE values

Moreover, I believe MAE would work better with this data set because it is relatively small, so to compensate for the possibility of outliers significantly impacting the performance of the model, MAE measures the average magnitude of the errors in the predictions made by the model. The lower the RMSE, the better the model is in predicting the target variable with greater accuracy and precision. According to the table below, we can see that the linear regression model has a lower MAE (80782) compared to the decision tree model (63491), indicating that it is better at predicting the outcome variable on average. The decision tree model has a higher RMSE (85228) compared to the linear regression model (284), indicating that it has higher variability in its predictions and is less accurate in predicting the outcome variable.

RMSE values for three models

Model 1: Linear Regression	Model 2: Decision Tree
284.2222590617779	85228.70750369267

Figure 6. Table of RMSE values

Given the analysis of a supervised regression model and a supervised classification model, linear regression is a better predictor and better fit for this data set. While the decision tree model may have a lower MAE, the higher RMSE suggests that it may not be the best model for this particular problem.

5. Appendix

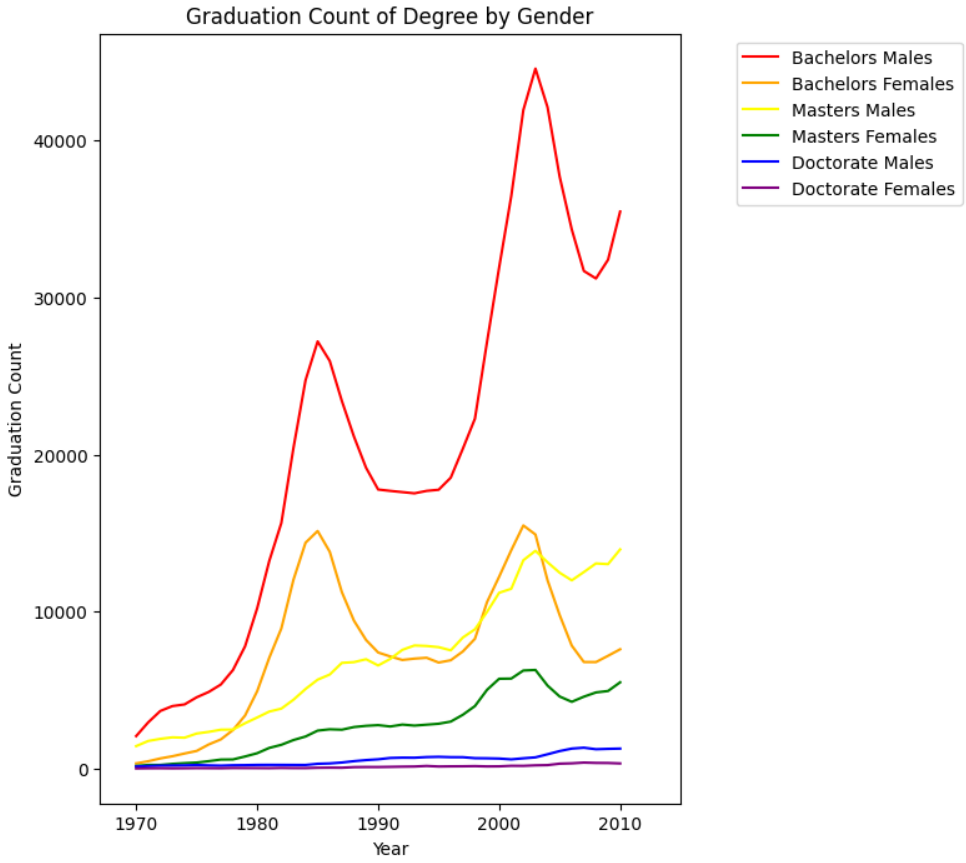

Figure 7. Time Series on Graduation Count of Varying Degree Levels

During the data exploration phase, one of the many data sets that piqued my interest captured information on various types of post secondary degrees conferred over 40 years. The figure above displays a time series with spikes in popularity of undergraduate degrees around 1985 and 2005. Further research could explore the social economic and historical context around these spikes in participation.

Given the scope of the times data, an interesting finding was noticing the change in naming conventions for college majors and more inclusive naming conversations for demographic data. For example, Before 1990s, "family and consumer sciences" used to be referred to as home economics. "security and protective services major", "R.O.T.C. and military sciences", "homeland securities" and "Military technologies" are all majors that used to exist in 1996, but no long appear in the roster of majors.

6. References

Digest of Education Statistics, 2021. National Center for Education Statistics (NCES) Home Page, a part of the U.S. Department of Education. (n.d.). https://nces.ed.gov/programs/digest/d21/tables/dt21$_3$22.10.*asp?current* = *yes*