Ranking Prediction of Higher Education Institutes using Financial and Expenditure Data

Mohamad Zeini Jahromi
Sep 18th, 2017

Proposal

Domain Background

The total number of awards, certificates, and degrees completed each year in post-secondary education institutions, have been widely used to evaluate their performances and are indication of relative success and ranking of these institutions throughout the nation. Many researches have studied the effects of different parameters such as financial aid, institutes funds, revenue, expenditures and etc. on their respective completion rates. The results of these studies would help institutions to decide how to allocate funds to their segments more effectively and create a productive money flow within their systems. On the other hand, the costs of higher education in the United Sates are high and having a knowledge of success rate and ranking of a specific institution is important to both students and their families for such an investment.

This project focuses on the studying the relationship between institutions financial aid, expenditure and completion rates (the total number of award, certificate, and degree completed). Furthermore, a predictive model will be developed in order to predict the completion rates using the financial aid and expenditure data of respective institutions.

This study is inspired by Udacity's capstone project guidelines and uses the IPEDS dataset (Integrated Postsecondary Education Data System Delta Cost Project Database), which is based on surveys on finance, enrollment, staffing, completions, and student aid of post-secondary education institutions within the US.

Problem Statement

As stated before, the completion rates (the total number of award, certificate, and degree completed) is a measure of success and ranking of institutions. In this project we are seeking to find out whether the completion rates could be predicted based on financial aid and expenditure of post-secondary education institutions either combined or individually and also which one of these parameters is more significant in terms of predicting the labels.

Datasets and Inputs

The dataset are from IPDES database (Integrated Postsecondary Education Data System Delta Cost Project Database), which is based on surveys on finance, enrollment, staffing, completions, and student aid of post-secondary education institutions in the US. It consists of many features including financial aid, revenue and expenditure of institutes across the US. The dataset for academic years of 2000 – 2012 has total number 87560 entries and 1007 features. All expenditure categories (around 20) will be used as the expenditure features and the same procedure will be followed for the financial aid categories. Also the categories that contain total completions value are going through further preprocessing to create the completion rates labels.

The selected features are numerical, and since the dataset contains a significant amount of null data (missing data or unreported data), we will perform data cleaning which results in eliminating some features and data points.

Solution Statement

Different algorithms (K-Nearest Neighbor Regressors, Support Vector Regressors, Random Forest Regressors and etc.) will be used to create predictive models. These models will be able to predict a specific institution completion rate based on its expenditure or financial aid features.

The data cleaning process will be performed on the dataset to eliminate outliers or missing values and then cross-validation is used to split the dataset into a training and a testing datasets.

The Principal Component Analysis will be implemented to find the dimensions that explain the most of the variance. The top components become predicting features for creating model. Before the PCA procedure, the standardizing and transforming procedures will be performed on the dataset. The models will be tuned using Grid Search function using sets of hyperparameters.

Benchmark Model

The initial model will be trained on the original dataset and assigned as the Benchmark Model. The next models will be trained using PCA-transformed datasets and will be tuned and optimized further. The evaluation metrics and performance scores (for example R-squared) will be used to compare the results of the benchmark model and other models.

Evaluation Metrics

We will use two different evaluation metrics to quantify the performance of both the benchmark model and the solution models. Mean Absolute Error and R-squared. Mean Absolute Error (MAE) calculates the average of the difference between predicted value and actual value. Root Mean Squared Error (RMSE) takes the root of the mean squared distance between predicted value and real value.

R-squared (R2) calculates the variance of the actual dataset and calculates the proportion for which that the predicted data can be accountable. The residual of the variance indicates the variance caused by the difference between actual data and predicted data. The much the variance being explained by predicted data, the higher the score is, which also indicates higher accuracy. R2, however, inflates when adding more predictors (variables) to the metric. The variance caused by predicted value hence increase even without model improvement. Adjusted R-squared score (Adj R2) is developed to counter the inflation and adding a penalty for additional variables entering the metric. Adj R2 is always smaller or equals to R2 score. Predicted R2 is another score that helps to determine if a model fits the original data but less capable of providing valid predictions on new data points.

When comparing MAE and RMSE, RMSE puts more weights on the larger error, and MAE behaves less sensitive to outliers. When comparing the model of original data and the model of the transformed dataset, it is likely the dimensionality changes. R2 score should not be the metric due to the inflation, and Adjusted R2 will be a better choice. While R2 ranges from 0 to 1, Adj R2 can be negative, which makes it more complicated than R2 to be understood. This feature makes Adj R2 rather easy to understand and straightforward.

Project Design

At the beginning, the dataset will be loaded with python pandas library, and will be stored as a data frame. After removing outliers, missing values, unnecessary data points and features, the cross-validation function will be used to split the data into training and the testing datasets (80% to training datasets and 20% to testing datasets).

The Principal Component Analysis will be implemented to find the dimensions that explain the most of the variance. The top components become predicting features for creating next models. Before the PCA procedure, the standardizing and transforming procedures will be performed on the dataset.

Different algorithms will be used to create predictive models. These models will be able to predict a specific institution completion rate based on its expenditure or financial aid features. The following are examples of these algorithms.

- K-Nearest Neighbor Regressor: The model estimates value by the closest data points, and the model costs less with minimal tuning.
- Decision Tree Regressor: The model is easy to implement and also performs very fast. Once trained the model predicts very quick. The model gives a clear structure of how it makes a prediction, but the downside is the prediction may lack interpretation.
- Support Vector Regressor: The model is also a commonly used model. SVR has solid founding theory, less prone to over-fitting, and need less tuning.
- Multiple Layer Perceptron Regressor: The model uses the structure of a neural network and is regaining popularity due to numerous applications on voice and image recognition.
- Random Forest Regressor: The model creates multiple decision trees and trains with data through bootstrap sampling. Every node on the branch will randomly choose a small amount of the features. The trees will be tested with data not sampled. This method can avoid over-fitting to training data by its randomness.

Then, the models will be optimized and tuned using Grid Search function and sets of hyperparameters. The metrics mentioned before, will be imported from scikit-learn package and will be used to evaluate different models and their testing results.

The final goal would be finding the most accurate model which is capable of predicting labels with highest performance scores. An important task when performing supervised learning on a dataset like this, is determining which features provide the most predictive power. By focusing on the relationship between only a few crucial features and the target label, we simplify our understanding of the phenomenon, which is most always a useful thing to do. Moreover, using the Principal Component Analysis, a list of expenditure and financial aid features will be selected to be used as inputs for the final model in order to speed up the prediction process while maintaining the accuracy in an acceptable range.

References

1. Integrated Postsecondary Education Data System Delta Cost Project Database, https://nces.ed.gov/ipeds/deltacostproject/
2. Rankings of universities in the United States, https://en.wikipedia.org/wiki/Rankings_of_universities_in_the_United_States
3. 2009–2010 College Rankings: National Universities, http://www.parchment.com/