Capstone Proposal

Thomas M Hughes
October 27, 2016

Domain Background

For this project, I am interested in producing accurate predictions of enrollment at American Universities. Universities require accurate enrollment numbers in order to properly allocate resources, from financial resources to human resources. However, historically, these predictions have proven notoriously difficult to make.

Historically, many universities use rather simple prediction models. They start by looking at their historical attrition rate, usually as an average, and subtract that from the target enrollment for the year. Then they add in the number of students who have been admitted for the upcoming year. Thus, their model is a simple linear model as such:

$y = mx + b$

With the following definitions:
y: enrollment prediction
m: current student population rate
x: historical attrition rate average
b: students admitted for this calendar year

A better model might be based on a variety of institutions attributes that may help predict more accurately, independent of the attrition rate.

Citations:

- http://sites.williams.edu/wpehe/files/2011/06/DP-26.pdf
- http://www.uwsp.edu/enrollmanage/Documents/Predicting%20Enrollments.pdf
- http://spu.edu/depts/idm/docs/publications/JW_Publication07.pdf

Problem Statement

This project is attempting to discover a model that can accurately predict student enrollment numbers at universities based on measures of institutional attributes. This is quantifiable as $y = f(x)$, where ‘y’ is the predicted student enrollment number, ‘x’ is a set of measures of institutional attributes, and ‘f’ is our model. ‘y’ can be measured by taking a count of enrolled students at an institution in a given year, as can the attributes associated with that institution. Furthermore, this problem recurs annually at every institution of higher education.

Datasets and Inputs

For this project, I will be using the Integrated Postsecondary Education Data System (IPEDS) Delta Cost Project Database from 2000-2012. This database includes information about higher education institutions, including finance, enrollment, staffing, completions, and student aid. It can be obtained at the following address: http://nces.ed.gov/ipeds/deltacostproject/download/IPEDS_Analytics_DCP_87_12_CSV.zip

IPEDS collects this data annually via surveys distributed to all post-secondary institutions in the United States that participates in federal student financial aid programs. (Every post-secondary institution with even a remotely good reputation participates in federal student financial aid programs. Many with poor reputations do as well.) This data set contains 974 attributes with 87,560 observations. Among these attributes, is a straight-forward 'enrollment' value, which will be our target variable. Other attributes of interest include tuition, endowment, number of employees, faculty salaries, federal grant data, and the like.

After some feature selection, to reduce the dimensionality, I intend to feed these inputs into a series of supervised regression algorithms, to see which produces the most accurate predictions, measured by squared error.

Solution Statement

A solution to this problem would accept a set of attributes about an institution (x), run them through a model (f), to produce a predicted enrollment number (y). A good solution would have predictions that have a low squared error value, when compared against a withheld test set.

Benchmark Model

Given that existing institutions tend to base their enrollment projections on a combination of attrition and new admissions, a benchmark model based on the retention rate and admission number should provide a good starting point. Specifically, a linear regression model that only takes the average of 'ftretention_rate' and 'ptretention_rate' plus the 'admitcount' to predict 'total_enrollment', with the mean squared error for a point of comparison. This should be close to the existing simple model used at institutions.

Evaluation Metrics

I will be using mean squared error as an evaluation metric that can be used to quantify the performance of both the benchmark and the solution model. The evaluation metric is derived by taking an array of predicted values, subtracting them from the observed values in the test set, squaring them, and taking the average over the whole set. Lower mean squared error is better.

For the mathematical representation, see:
https://en.wikipedia.org/wiki/Mean_squared_error#Predictor

Project Design

I plan to use a three-stage workflow to find a solution to this problem.

Given that the data set has a very large number of attributes to work with, Stage 1 will require some preprocessing. In particular, two strategies will be employed to deal with missing data. For the target variable, missing values will be dropped. For the non-target attributes, I plan to generate two different data sets, to see how they perform. The first will simply drop cases missing values for important attributes, resulting in an overall lower sample size. The second will impute the missing values by taking the median value for that attribute. Both sets will be used independently for each model to see which performs better. I will then drop outliers along the target variable dimension. Once the data has been cleaned, I will create the baseline model described above.

Stage 2 will primarily focus on generating models that (hopefully) perform better than the baseline. I will begin with feature selection, removing features from the data that is redundant (such as other partial measures of enrollment), or provides no useful information (such as school name). With the remaining potential features, I will perform Independent Component Analysis (ICA) to reduce the number of dimensions to a number that is viable, given the size of our data after the cleaning from stage 1, with the curse of dimensionality in mind. I will then generate train/test splits for the data sets, probably with 80/20 proportions, though that may vary based on the size of the datasets. With the data setup, I will then generate a series of models for comparison, using a variety of regression models (Linear Regression, Support Vector Regression, and Random Forest Regression). The Mean Squared Error of each of these models will be compared against the baseline model, and visualizations of each predicted value vs. observed value in the test set will be generated and presented.

In stage 3, I plan to optimize the model that performed best from stage 2. I plan to do this with a parameter grid search to generate optimum parameters for the model. A final comparison of how the optimized model performs against the baseline model will then be presented.

Simplified Workflow Outline:

Stage 1: Initial Date Setup and Create Baseline Model

- Import data
- Clean Missing Values
    - Drop missing values
    - Impute missing values.
- Drop Outliers
- Create baseline model (described above)

Stage 2: Dimensionality Reduction and Model Comparisons

- Remove Features that provide no information or are redundant
- Perform Independent Component Analysis to reduce dimensionality
- Create Train Test Split
- General Multiple Potential Models for Comparison
    - LinearRegression model
    - Support Vector Regression model
    - RandomForest Regression ensemble model
- Generate visual plots of each of the 4 models, to compare performance
    - Compare MSE values of the 4 models
    - Note: Watch for dangerous overfitting

Stage 3: Optimization:

- Pick the best model of the 4, and perform a grid search to find optimal parameters
- Report final comparison of optimized model to baseline model