Estimating University Enrollments

Machine Learning Engineer Nanodegree Capstone Project

Thomas M Hughes

10/28/16

Definition

Project Overview

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. 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.

As someone who has previously worked at institutions of higher education, I am particularly interested in using Machine Learning techniques to find efficiency gains and cost-saving opportunities like this one.

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 215,613 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.

In this section, look to provide a high-level overview of the project in layman's terms. Questions to ask yourself when writing this section:

- Has an overview of the project been provided, such as the problem domain, project origin, and related datasets or input data?
- Has enough background information been given so that an uninformed reader would understand the problem domain and following problem statement?

Rubric: Student provides a high-level overview of the project in layman's terms. Background information such as the problem domain, the project origin, and related data sets or input data is given.

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.

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.

To find a solution, 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).

In this section, you will want to clearly define the problem that you are trying to solve, including the strategy (outline of tasks) you will use to achieve the desired solution. You should also thoroughly discuss what the intended solution will be for this problem. Questions to ask yourself when writing this section:

- Is the problem statement clearly defined? Will the reader understand what you are expecting to solve?
- Have you thoroughly discussed how you will attempt to solve the problem?

- Is an anticipated solution clearly defined? Will the reader understand what results you are looking for?

Rubric: The problem which needs to be solved is clearly defined. A strategy for solving the problem, including discussion of the expected solution, has been made.

Metrics

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. The Mean Squared Error of each candidate model (described above) 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 this section, you will need to clearly define the metrics or calculations you will use to measure performance of a model or result in your project. These calculations and metrics should be justified based on the characteristics of the problem and problem domain. Questions to ask yourself when writing this section:

- Are the metrics you've chosen to measure the performance of your models clearly discussed and defined?
- Have you provided reasonable justification for the metrics chosen based on the problem and solution?

Rubric: Metrics used to measure performance of a model or result are clearly defined. Metrics are justified based on the characteristics of the problem.

Analysis

Data Exploration

The IPEDS Delta Cost Project Database provides survey results from American institutions of higher education that accept federal financial aid funds during the years of 1987 to 2012. This dataset provides us with 215,613 observations, across 974

attributes. Because this database is split into two separate files (1987-99, 2000-12), we will need to concatenate the two files together into a single dataframe. Once we isolate our target variable ('total_enrollment'), we can obtain the following summary statistics:

Table: Enrollment Summary Statistics

count	153,168
mean	2,809.486148
standard deviation	7,540.511589
minimum	0
25%	101
50%	481
75%	2344
max	380,232

From these summary statistics, it appears that there will be a rather significant issue with missing data (our count is lower than our original observations), and furthermore, it looks like there will be a significant issue with outliers (the max and minimum are very far from the 50% mark). Also, with 974 features to work with, quite a bit of feature selection is going to need to be done to avoid problems with the curse of dimensionality. Furthermore, some of the features are categorical, but potentially useful, like 'zip', 'census_region', 'state', etc.

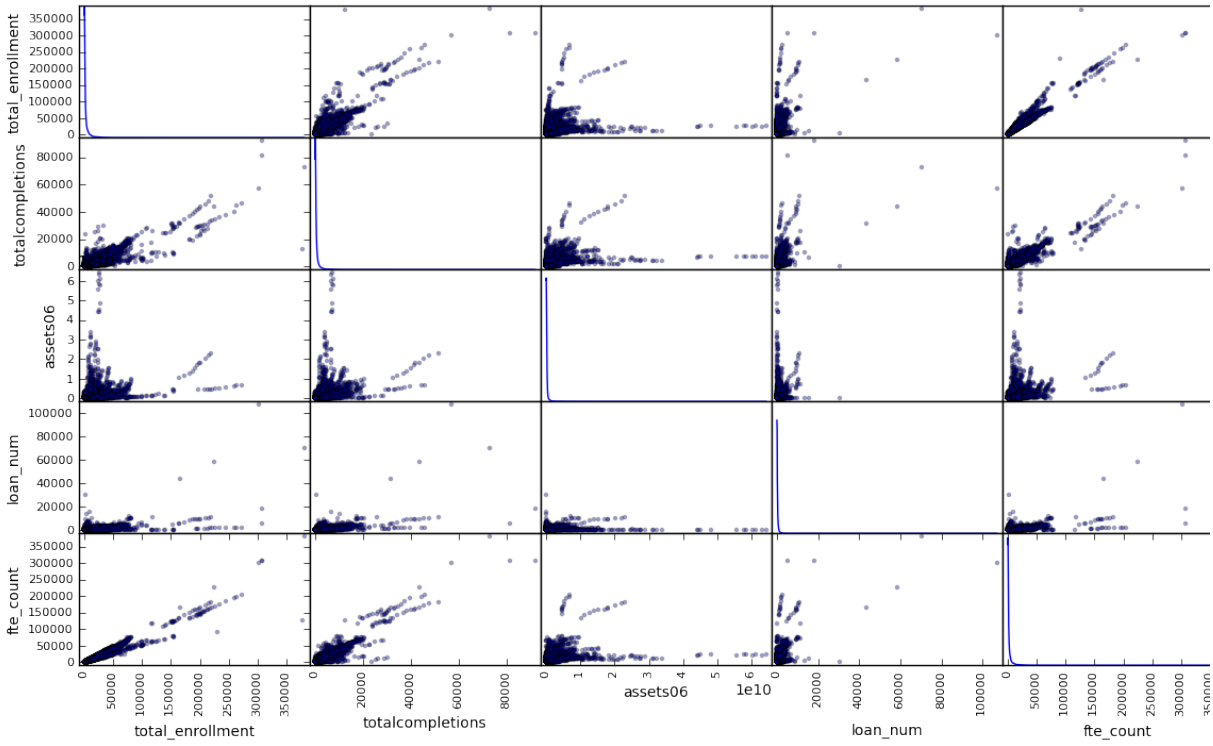
However, most of the features present boolean, continuous, or ratio values. Some of the features are composites of other features, such as 'total_enrollment' being a combination of 'total_fulltime' and 'total_parttime' (among others). A combination of feature selection and component analysis will be necessary to reduce the feature space.

In this section, you will be expected to analyze the data you are using for the problem. This data can either be in the form of a dataset (or datasets), input data (or input files), or even an environment. The type of data should be thoroughly described and, if possible, have basic statistics and information presented (such as discussion of input features or defining characteristics about the input or environment). Any abnormalities or interesting qualities about the data that may need to be addressed have been identified (such as features that need to be transformed or the possibility of outliers). Questions to ask yourself when writing this section:

- If a dataset is present for this problem, have you thoroughly discussed certain features about the dataset? Has a data sample been provided to the reader?
- If a dataset is present for this problem, are statistics about the dataset calculated and reported? Have any relevant results from this calculation been discussed?
- If a dataset is not present for this problem, has discussion been made about the input space or input data for your problem?
- Are there any abnormalities or characteristics about the input space or dataset that need to be addressed? (categorical variables, missing values, outliers, etc.)

Rubric: If a dataset is present, features and calculated statistics relevant to the problem have been reported and discussed, along with a sampling of the data. In lieu of a dataset, a thorough description of the input space or input data has been made. Abnormalities or characteristics about the data or input that need to be addressed have been identified.

Exploratory Visualization



Since this data set includes such a large feature space (close to a thousand features), attempts to sample some of the data and the feature behavior will be done randomly to try to get a rough sense of how the data looks. When we take a random sample of features, and plot them as pairs, we notice that the data is definitely not normally

distributed, and there are quite a few outliers present. However, the scatter plotting of random samplings of pairs does seem to indicate there may be some linear relationships in the data that we may be able to capture later. Some normalization will also likely have to occur.

In this section, you will need to provide some form of visualization that summarizes or extracts a relevant characteristic or feature about the data. The visualization should adequately support the data being used. Discuss why this visualization was chosen and how it is relevant. Questions to ask yourself when writing this section:

- Have you visualized a relevant characteristic or feature about the dataset or input data?
- Is the visualization thoroughly analyzed and discussed?
- If a plot is provided, are the axes, title, and datum clearly defined?

Rubric: A visualization has been provided that summarizes or extracts a relevant characteristic or feature about the dataset or input data with thorough discussion. Visual cues are clearly defined.

Algorithms and Techniques

Start

In this section, you will need to discuss the algorithms and techniques you intend to use for solving the problem. You should justify the use of each one based on the characteristics of the problem and the problem domain. Questions to ask yourself when writing this section:

- Are the algorithms you will use, including any default variables/parameters in the project clearly defined?
- Are the techniques to be used thoroughly discussed and justified?
- Is it made clear how the input data or datasets will be handled by the algorithms and techniques chosen?

Rubric: Algorithms and techniques used in the project are thoroughly discussed and properly justified based on the characteristics of the problem.

Benchmark

Start

In this section, you will need to provide a clearly defined benchmark result or threshold for comparing across performances obtained by your solution. The reasoning behind the benchmark (in the case where it is not an established result) should be discussed. Questions to ask yourself when writing this section:

- Has some result or value been provided that acts as a benchmark for measuring performance?
- Is it clear how this result or value was obtained (whether by data or by hypothesis)?

Rubric: Student clearly defines a benchmark result or threshold for comparing performances of solutions obtained.

Methodology

Data Preprocessing

Start

In this section, all of your preprocessing steps will need to be clearly documented, if any were necessary. From the previous section, any of the abnormalities or characteristics that you identified about the dataset will be addressed and corrected here. Questions to ask yourself when writing this section:

- If the algorithms chosen require preprocessing steps like feature selection or feature transformations, have they been properly documented?
- Based on the Data Exploration section, if there were abnormalities or characteristics that needed to be addressed, have they been properly corrected?
- If no preprocessing is needed, has it been made clear why?

Rubric: All preprocessing steps have been clearly documented. Abnormalities or characteristics about the data or input that needed to be addressed have been corrected. If no data preprocessing is necessary, it has been clearly justified.

Implementation

Start

In this section, the process for which metrics, algorithms, and techniques that you implemented for the given data will need to be clearly documented. It should be abundantly clear how the implementation was carried out, and discussion should be made regarding any complications that occurred during this process. Questions to ask yourself when writing this section:

- Is it made clear how the algorithms and techniques were implemented with the given datasets or input data?
- Were there any complications with the original metrics or techniques that required changing prior to acquiring a solution?
- Was there any part of the coding process (e.g., writing complicated functions) that should be documented?

Rubric: The process for which metrics, algorithms, and techniques were implemented with the given datasets or input data has been thoroughly documented. Complications that occurred during the coding process are discussed.

Refinement

Start

In this section, you will need to discuss the process of improvement you made upon the algorithms and techniques you used in your implementation. For example, adjusting parameters for certain models to acquire improved solutions would fall under the refinement category. Your initial and final solutions should be reported, as well as any significant intermediate results as necessary. Questions to ask yourself when writing this section:

- Has an initial solution been found and clearly reported?
- Is the process of improvement clearly documented, such as what techniques were used?
- Are intermediate and final solutions clearly reported as the process is improved?

Rubric: The process of improving upon the algorithms and techniques used is clearly documented. Both the initial and final solutions are reported, along with intermediate solutions, if necessary.

Results

Model Evaluation and Validation

Start

In this section, the final model and any supporting qualities should be evaluated in detail. It should be clear how the final model was derived and why this model was chosen. In addition, some type of analysis should be used to validate the robustness of this model and its solution, such as manipulating the input data or environment to see how the model's solution is affected (this is called sensitivity analysis). Questions to ask yourself when writing this section:

- Is the final model reasonable and aligning with solution expectations? Are the final parameters of the model appropriate?
- Has the final model been tested with various inputs to evaluate whether the model generalizes well to unseen data?
- Is the model robust enough for the problem? Do small perturbations (changes) in training data or the input space greatly affect the results?
- Can results found from the model be trusted?

Rubric: The final model's qualities — such as parameters — are evaluated in detail. Some type of analysis is used to validate the robustness of the model's solution.

Justification

Start

In this section, your model's final solution and its results should be compared to the benchmark you established earlier in the project using some type of statistical analysis. You should also justify whether these results and the solution are significant enough to have solved the problem posed in the project. Questions to ask yourself when writing this section:

- Are the final results found stronger than the benchmark result reported earlier?
- Have you thoroughly analyzed and discussed the final solution?
- Is the final solution significant enough to have solved the problem?

Rubric: The final results are compared to the benchmark result or threshold with some type of statistical analysis. Justification is made as to whether the final model and solution is significant enough to have adequately solved the problem.

Conclusion

Free-Form Visualization

Start

In this section, you will need to provide some form of visualization that emphasizes an important quality about the project. It is much more free-form, but should reasonably support a significant result or characteristic about the problem that you want to discuss. Questions to ask yourself when writing this section:

- Have you visualized a relevant or important quality about the problem, dataset, input data, or results?
- Is the visualization thoroughly analyzed and discussed?
- If a plot is provided, are the axes, title, and datum clearly defined?

Rubric: A visualization has been provided that emphasizes an important quality about the project with thorough discussion. Visual cues are clearly defined.

Reflection

Start

In this section, you will summarize the entire end-to-end problem solution and discuss one or two particular aspects of the project you found interesting or difficult. You are expected to reflect on the project as a whole to show that you have a firm understanding of the entire process employed in your work. Questions to ask yourself when writing this section:

- Have you thoroughly summarized the entire process you used for this project?
- Were there any interesting aspects of the project?
- Were there any difficult aspects of the project?
- Does the final model and solution fit your expectations for the problem, and should it be used in a general setting to solve these types of problems?

Rubric: Student adequately summarizes the end-to-end problem solution and discusses one or two particular aspects of the project they found interesting or difficult.

Improvement

Start

In this section, you will need to provide discussion as to how one aspect of the implementation you designed could be improved. As an example, consider ways your implementation can be made more general, and what would need to be modified. You do not need to make this improvement, but the potential solutions resulting from these changes are considered and compared/contrasted to your current solution. Questions to ask yourself when writing this section:

- Are there further improvements that could be made on the algorithms or techniques you used in this project?
- Were there algorithms or techniques you researched that you did not know how to implement, but would consider using if you knew how?
- If you used your final solution as the new benchmark, do you think an even better solution exists?

Rubric: Discussion is made as to how one aspect of the implementation could be improved. Potential solutions resulting from these improvements are considered and compared/contrasted to the current solution.