Estimating University Enrollments

Machine Learning Engineer Nanodegree Capstone Project

Thomas M Hughes

11/2/16

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

[[Exploratory Visual]]

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

Since the data set includes the target variable of enrollment, this will be a supervised learning problem. Furthermore, since the target variable is a number and not a category, this qualifies as a regression problem. Because of that, I will be focusing on regression algorithms and techniques.

The first thing I want to note is that, with the exception of the benchmark model, I plan to use Independent Component Analysis in all my candidate models to reduce the feature space. I specifically chose ICA for this feature reduction because the data exploration suggests that many of the columns in the data are different measures of the same underlying attribute, and each of those measures seem to be capturing some cross noise from the other attributes. ICA should help to capture the independent signals that the measures seem to be indicating.

Furthermore, with ICA, it will be possible to use many more independent features than the benchmark does, as the benchmark really only uses two features (admission count and retention rate). By contrast, the ICA feature reduction will allow the candidate models to use up to 17 independent component features.

In addition to the ICA done on the feature space, I will be examining three candidate regression models. Since the current technique predominantly used to predict enrollments is a linear regression model with few features, a good candidate would be another linear regression model using the ICA features. That is, perhaps better results could be obtained by using the same model but with more (and perhaps better) features. Aside from this difference, both will be using the default parameters

As a separate condition, perhaps the underlying data is linearly separable. Considering the high degree of dimensionality, it is hard to tell this in advance. However, if it turns out that the data is linearly separable, a support vector approach to the problem could be promising. As such, Candidate #2 will be a Support Vector Regressor, using the default settings.

Finally, ensemble methods have had a high degree of success with a number of machine learning problems. Thus, my third candidate will be a Random Forest Regressor. That is, it will be a regressor that will take multiple Decision Tree Regressors on the data, and average the result. Again, this will be done using the default parameters.

After all four models have been tested, I will take the best performing model and run a Grid Search to attempt to find the optimal parameters, and better tune the model.

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

Existing institutions tend to base their enrollment projections on a combination of attrition and new admissions,[1] a benchmark model based on the retention rate and admission number should provide a good starting point. Specifically, a linear regression model 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.

It should be noted, I cannot get the exact same model as is usually used, as I do not have that full data available. Most institutions have a precise count of how many are re-enrolling, and they also have an institutional target goal for enrollment. With these two pieces, institutions actively aim for an admit count that will produce their target enrollment. As a result, I am using the retention rate as a proxy for a re-enrollment measure, and the admit count to give an estimate of the target enrollment for the institution.

Mean squared error (MSE) is a good metric for comparison here, as it captures how close the model predictions are to the actual observed enrollments, with a penalty for being further away. The benchmark model has a MSE of approximately 2.913

(normalized). This is not a great model, but it does provide a starting point to evaluate from.

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

A significant amount of data preprocessing will be required, based on what was seen in the data exploration section. I proceed with the following procedures: minimal feature selection, drop missing target values, handling outliers, imputing missing values, scaling and normalizing the data, splitting the data for cross validation, and finally, Independent Component Analysis (ICA) to reduce the feature space.

Feature selection was fairly minimal. All columns that contained strings were dropped, as these columns either contained no relevant information (such as institution name), or they contained data that was implied in other columns. For example, the string column of state names can more or less be deduced by the zip code column. Since the strings are harder to work with, and provide no clear advantage, those columns are dropped from the data. I also made sure to convert the remaining values to numerics, as there were some cells that imported invisible characters, making the values appear as non-numbers.

Next, I dropped all rows that were missing the target value of 'enrollment'. While it is worth attempting to impute missing values for the other columns, my worry is that imputing values for the target variable may end up distorting the results of the model.

This reduces the number of observations we're working with from 215,613 down to 153,168. This is still quite a few observations, so there is no worry of a small n problem.

In the data exploration section, the visualizations revealed some significant outliers in the data that may produce odd results. Furthermore, since the plan is to impute the missing data, large outliers can have a significant impact on the results. So, at this point, I dropped any rows of the target variable which were greater or less than 1.5 times the interquartile range. I limited this outlier removal to only the target variable, on the assumption that the largest and smallest institutions would be the least normal along multiple dimensions.

After outliers were removed, I then imputed the missing data from the non-target variables. I should mention, my first attempt was not to impute the missing data, but simply drop any row that had any missing data. This resulted in zero observations. More selective dropping might be possible, but with over 150,000 observations and close to 1,000 feature dimensions, manually figuring out what could and could not be dropped was impractical. Imputing the missing values provided an alternative. I did this first by using the pandas DataFrame interpolate method, which would attempt to sort the values in a column, and would estimate that the missing values between actual observations followed a linear pattern. For values that could not be interpolated, the column mean was used. These are not perfect estimates, but since they are not done to the target values, they should not artificially inflate the predictive value of the resulting models.

[[[Image of Imputed Data]]]

As we see above, at this state the pre-processed data appeared to have some issues with skew. So I wanted to normalize the data so that it would have a roughly gaussian distribution, since many models assume that. To do this, I scaled the data using a min-max scale. This removed any negative values. I then multiplied that by 10,000 to avoid potential problems with really tiny numbers, and finally I took the natural logarithm of the result. Furthermore, any values that resulted in negative infinity, as the result of taking the logarithm of zero, I converted those values back to zero.

[[[Image of normalized data]]]

As we see in this image, the resulting data looks more normally distributed, but we can also spot a couple of irregularities. In many of the images, there are observations that appear extremely regular, almost like perfectly horizontal or vertical lines. Those points are artifacts of the imputing process. That is, the estimated values end up looking a little too perfect in the data. Again, since the imputing process was not done to the target variable, only the inputs, this artifact should not artificially improve the result of our model. Worst case scenario, the imputing artifacts make our model have a worse MSE score. However, if the estimates are good enough, they may make the model more accurate in reality.

After all this was done, the data was then split into 80% training set, 20% testing set, for cross validation purposes.

Finally, since the data set contained such a large number of features, I needed to dramatically reduce the feature space to deal with the issue of the curse of dimensionality. To do this, I used Independent Component Analysis (ICA) to transform the inputs down to 17 independent dimensions. I used ICA because the feature space often had several columns that seemed to be taking measurements of similar values (such as full time and part time admissions, or in-state and out-of-state admission, etc), but often these measures would have noise from some other hidden feature as well. ICA should help to isolate out the actual independent features behind the measures. I chose 17, because we have sufficient observations to handle up to 17 features without running into problems with the curse of dimensionality.

Once all this has been done, I have accounted for the missing values, the skew in the distribution of the data, and excess of feature space in light of the curse of dimensionality. The data is now ready for analysis.

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

By contrast to the data preprocessing, the model implementation was very straight forward. Four regression models were tested: the benchmark model (linear regression with 3 non-ICA features), linear regression with ICA, Support Vector Regressor (SVR) with ICA, and Random Forest Regressor with ICA.

For each candidate, the model was trained with the training inputs and the training enrollment targets. The model was then used to make predictions for the test inputs. Finally, I generated two sets of Mean Squared Error results, for both the training data and the testing data, by using the predicted values from the inputs compared to the actual observed values of enrollment. The MSE for the training data was compared to the test to get a sense model overfit, while the MSE test value is our primary metric for comparing usefulness of each candidate model. No expected complications or changes arose in the implementation.

Model	**Training Normalized MSE**	**Testing Normalized MSE**
Benchmark (Linear no ICA)	2.92670403437	2.9127528257
Candidate 1: Linear w/ ICA	0.572164100866	0.568313314229
Candidate 2: SVR w/ ICA	2.73706191612	2.71240878953
Candidate 3: Random Forest w/ ICA	0.0683715696409	0.357933570385

As can be seen in the summary results table, the Benchmark model did not perform particularly well, but neither did the Support Vector Regressor. Furthermore, the SVR model ran extraordinarily slow, while producing results almost as bad as the benchmark. This makes it easy to dismiss the SVR as not providing much value over the benchmark model.

Choosing between candidate 1 and candidate 3 is a bit more tricky. Both perform significantly better than the benchmark model. The Random Forest model ultimately has the lower Mean Squared Error on the testing data, but the MSE on the training data is worrying. There does appear to be a real danger of overfitting with candidate 3, where candidate 1 does not appear to have the same problem. Having said that, even with the overfitting on the training data, candidate 3 still performs much better than any other candidate on the test data.

Thus, the Random Forest Regressor with ICA is our initial solution to the problem of enrollment prediction.

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

With the Random Forest selected as the final model for the solution, I performed a Grid Search to find optimal parameters. Specifically, I had the Grid Search test various numbers of trees developed for the forest (5, 10, 20 or 40; default is 10). I also had the grid search experiment with using fewer features (auto, sqrt, or log2; default auto=17).

After the grid search was performed, the best and final parameters resulted in a testing normalized MSE of 0.317, which is better than the untuned model. Unfortunately, the normalized MSE of the training data was 0.048, suggesting there is still a real danger of overfit with this model.

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

The final model appears to be a reasonable solution to the problem of estimating enrollment at institutions of higher education. By using a wide variety of institutional attributes related to the student body makeup, the faculty and staff makeup, funding sources and the like, give us a reasonable way to estimate current enrollment. This is true even if some of the institutional attributes need to be crudely estimated. By running Independent Component Analysis, it becomes possible to isolate distinct signals from the attributes, which can then be run through a Regressor. A Random Forest Regressor — running multiple Decision Tree Regressors and bagging the results — produces the best final estimate.

To get the true estimate of enrollment, a little bit more work needs to be done. Remember, the model predicts a transformed estimate of enrollment that has been normalized. To get the actual estimate, the value produced by the model must be used as an exponent for *e* to get the reverse of the natural logarithm, and then divided by the enlarging factor that was used to avoid small numbers (10,000), and then de-normalized by multiplying by the original non-outlier maximum enrollment, minus the original non-outlier minimum enrollment, plus the non-outlier minimum enrollment.

Because the model was tested using cross validation, there is reasonable confidence that the model can produce reasonably accurate results, with most institutions. The estimates are, on average, only off by about 1 student on the unseen testing data. If this sounds unreasonably accurate, it is worth remembering that the median institution — once outliers are removed — only has 314 students enrolled. This is true because the huge outlier institutions have been removed from the data set. Less accuracy would likely be found if the huge outlier institutions were re-introduced, or if attempts were made to predict on the outlier data.

There is one final concern about trusting the model: there remain signs of overfit, mentioned previously. However, given that the Random Forest does not perform absurdly better than the Linear Regression Model — which does not show signs of overfit — there is reason to think that the results of the final model are a plausible fit for the data.

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

To summarize, the Random Forest Regressor with ICA does outperform the benchmark established earlier, when comparing Mean Squared Errors of predictions on the test data. The difference is significant enough to suggest that the model presented here is enough of an improvement to consider it the new benchmark. I would not suggest the problem is solved, despite these improvements. In my conclusion, I will discuss possible improvements I think can be made to the model.

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

Working with a data set that provides such a large number of observations and such a large number of features makes it a bit difficult to visualize. However, I want to show at least one visual comparison that gives a sense of how well the final model tends to predict the target variable.

[[[Observations vs. Feature #5]]]

In this scatterplot, we see all of the actual observed enrollment values plotted against one of the results of the Independent Component Analysis, #5. #5 was selected because the final model registers it as having the highest feature importance when making predictions. The general shape of the plot is a bit odd. In general, as Feature #5 increases, so does enrollment, but there is a huge amount of variance. There is also a peculiar dip as feature #5 approaches zero from the negative side.

[[[Predictions vs. Feature #5]]]

This graph, with the green plot, is the same measures (total_enrollment plotted against Feature #5), but these are the predicted values, not the observed ones. Not only is the general shape the same, the predictor actually has the dip that the actual observations display as Feature #5 approaches zero from the negative side. This gives a pretty good visual sense in two dimensional space as to how the shape of the predictions appears pretty close to the shape of the actual observations.

If the reader is curious as to what Feature #5 actually is, that is a lot harder to explain. As mentioned above in the Data Preprocessing, ICA was used to reduce the 900+ features down to 17. The specific makeup of #5 can be seen in the following visualization of the feature weights:

[[[#5 Feature Weights]]]

As can be seen above, #5 combines signals from a large number of the original features, and especially weights some of the earlier features. Individually picking these out would take an extraordinary amount of time.

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

This project attempted to solve the problem of estimating enrollments at institutions of higher education. Using the PEDS Delta Cost Project Database, it is possible to evaluate a large number of institutional attributes that may help in predicting total enrollment. To do this, extensive preprocessing needed to be done to deal with outliers, missing data, and reduce the feature space. Once the data was prepared, four regression models were tested, comparing their Mean Squared Error on their predictions. Finally, the best performing model, the Random Forest Regressor, was run through a Grid Search to optimize its parameters.

The data preprocessing was by far the hardest part of this process. By having so many observations and so many feature columns, it was basically impossible to eyeball anything. It required taking a much more systematic approach to figure out how to make the data useful. The missing values on the inputs, in particular, was the hardest to work around. I am quite happy that the imputed estimates did not seem to hurt the analysis at all.

I think the final model works quite well as a solution to the problem for average institutions of higher education. The error rate ends up being quite low, and the predictions look to take a very similar shape to what is actually observed.

Perhaps unfortunately, most of the time what we ordinarily think of as average institutions are much larger than what this analysis actually ends up looking at. That is, the big well known schools are actually outliers, so this solution probably would not work well for institutions that are best known, or cater to the most students.

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

More careful — and painstakingly slow — feature selection would probably help the analysis here. Right now, the model takes almost all of the feature space into account, to some extent or another. Manually going through and removing features that appear highly correlated with one another — and appear to be measures of the same thing — might reduce the dangers of overfitting.

On consideration, perhaps the more important improvement for this analysis would be to be a *less* general model rather than a more general one. In particular, given that most students tend to be in larger institutions that were treated as outliers here, it may be more interesting to only do the analysis on the top quartile of institutions. Since most institutions are quite small, the solution here works for most schools, but not for the schools that contain the most students.

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.

---

[1] See: • 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