**MEMORANDUM**

**TO:** Professor Jingjing Li

**FROM:** Ellie McKinnon, Alex Struck, Bill Su, Grayson Thornton

**DATE:** 11/16/15

**SUBJECT:** Demystifying the Ivory Tower: Midterm Report

**Executive Summary**

Our project aims to help college applicants identify four-year undergraduate universities whose graduates generate the highest return on investment. We will make personalized college recommendations to users based on various attributes of the universities. Our final product will consist of a data tool where individuals can enter personal information and receive a list of schools that will generate the highest return. For instance, users will provide information such as family income level, region preference, and family legacy, and the data tool will return schools ranked by ROI.  
  
Our analysis will help users make more economically informed decisions about where to attend college. In doing so, we aim to help students tackle an extremely important personal decision, while simultaneously shedding light on the pressing societal issue surrounding college tuition and student debt. By providing students with the ROIs of different institutions, we hope to empower our users to lessen the economic pressure on both students and their parents and maximize the financial outcomes of their college decision.

**Business Understanding**

The primary goal of this project is to determine the best ROI for education given a college applicant’s preferences. This model will help a student choose the college with the best return on his investment in the future given his economic background and desired college attributes. The costs associated with a poor ROI college choice can be years of student debt, and financial distress. This model will be a success if it predicts the best college for a student with simplicity so that the output is easy to interpret with reasonable accuracy.

**Data Understanding**

For this project, we are going to use the CollegeScorecard data provided by the U.S. Department of Education. This provides college scorecard raw data from 1996 to 2013. These data sets are a collection of information from 3,846 universities across the United States.  Overall, the dataset records all attributes of college institutions including, but not limited to, its demography, geographic location, and future earnings for students 6 and 10 years after graduation. There are 124,699 entries in the dataset, with 1,731 variables. The dataset records data collected from 1996 to 2013. We have decided to use the 2011 data for our analysis due to abundance of information available.

We initially selected 203 variables out of the original dataset to explore so that our analysis is based only on relevant variables. From that data set we further paired down to 10 predictive attributes that are most important for students making the college decision. These attributes are: region, locale, midpoint of SAT scores (critical reading), midpoint of SAT scores (math), admissions rate, size of undergraduates, graduation rate, full-time faculty, median debt, average household income, and the target attribute is ROI. Summary statistics for appropriate attributes can be found in Appendix 6.

In order to calculate ROI, we have decided to divide 6 year annual income after graduation during 2011 by the cost of education for the same year. We decided to use cost of education for 2011 instead of 2005 due to the lack of availability for cost of education in the 2005 dataset. We consider the cost of education in 2011 as a simplified representation of the cost of education for an institution assuming that the relative cost of education of an institution does not change significantly in a 6-year interval. Numerical results of those calculations are then converted into categorical variables according to their specific value in which “low” represent negative ROI, “medium” represent ROI between 1 and 2, and “high” represent ROI that is higher than 2.

**Data Preparation**

Our biggest challenge when processing the dataset is the existence of missing values. Even with the 2011 dataset, which has the most information available, more than half of the variables are either not available or have missing values for over half of the instances. We have tackled this challenge in two ways.

For variables that can be accurately approximated from past year’s data, we have pulled data from other years. For example, the variable “LOCALE”, which describes the setting of an institution (urban, suburb, etc.), is not available for the year 2011. However, this variable is available for 2012. Because a college’s “LOCALE” is not likely to change in two years, we have decided to use 2012’s information instead (data processing done via Python).

For variables that are not available and cannot be accurately approximated, we have decided to use available data during the same year to approximate those variables. For example, “average household income” is not available for the year 2011, but similar variables such as “average household income for independent students” and “number of independent students” are available. We therefore approximated average household income for each school by summing the total household income for independent and dependent students and dividing by the sum of the population of those two student groups.

After conducting those two steps, we have also removed all attributes that have more than 6 missing values as they will provide very little predictive insights to our model.

However, despite our best efforts in filtering out missing values, variables such as SAT score and Admission rates are missing for a large portion of the institutions. Due to the importance of those two variables, we have decided to impute all missing variables by the global average of each attribute.

**Modeling**

Before conducting more detailed predictive analytics, we have first conducted several descriptive analytics on our dataset. Out of all instances of institutions in the final dataset, the majority of them has a medium return on investment. Only around 30% of the institutions have high return on investment. Additionally,  20% of the institutions have low (negative) return on investment, illustrating the fact that getting a college education usually provides better long-term benefits for students (Appendix 3).

In order to identify the most appropriate model, we outlined a set of objectives that our chosen model must satisfy and used an intelligent experimentation process to optimize accuracy and parameters. In doing so, we found that at this point in the modeling process, Random Forest best accomplishes these goals. While we began using Decision Tree and Rule Induction models and also experimented with Neural Net, SVM, and k-NN, Random Forest has been most effective thus far.

We believe that the best model for our data set and given business problem should be highly interpretable so that we can recognize the implications of the model and outline clear recommendations for our analytics consumers (students, teachers, and family members). Furthermore, the model must be scalable to large data sets in order to take in many attributes and account for a large set of colleges. Additionally, Random Forest is flexible and can account for multiple data types, which is important when dealing with many different university attributes. Finally, as opposed to using a Decision Tree or Rule Induction model, Random Forest helps mitigate risk by generating an ensemble of Decision Trees generated from subsets of the sample.

The following table outlines the Random Forest parameters used to achieve an accuracy of 69.18%:

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| **Number of Trees** | 50 |
| **Criterion** | Information Gain |
| **Minimal Size for Split** | 4 |
| **Minimal Leaf Size** | 2 |
| **Minimal Gain** | .01 |
| **Maximal Depth** | 20 |
| **Confidence** | .45000001 |
| **Subset Ratio** | .95 |

We arrived at these parameters through a combination of intelligent experimentation and parameter optimization. Our original Random Forest model output an accuracy of 52.27%, so we were able to increase accuracy by nearly 17% over multiple iterations. We found that the subset ratio, minimal gain, and tree size levers were particularly instrumental in driving higher accuracy. Moreover, manually determining ideal ranges and then using the optimization parameter within these ranges allowed us to both increase model accuracy and be more efficient in our modeling process.

Our final output is as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Actual High** | **Actual Low** | **Actual Medium** | **Precision** |
| **Pred. High** | 166 | 3 | 58 | 73.13% |
| **Pred. Low** | 1 | 80 | 28 | 73.39% |
| **Pred. Medium** | 67 | 80 | 286 | 66.05% |
| **Recall** | 70.94% | 49.08% | 76.88% |  |

This output shows that we can improve our current model’s ability to predict Low ROI colleges. In order to control for this, we will do further analysis around balanced sampling techniques and using the metacost operator to weight certain attribute values and increase recall and precision.

Further modeling efforts will center around ensemble methods and sampling techniques that may allow us to improve upon our current accuracy. For example, we hypothesize that grouping models, such as Random Forest, Rule Induction, and SVM, with a Voting Ensemble method could allow us to achieve a higher accuracy. Furthermore, using Bagging or Bootstrap ensemble methods will allow us to take the best features from many different models and combine them to create a more comprehensive model. While such efforts may be effective in increasing accuracy, we have to keep in mind our original objectives of developing a model that is highly interpretable, scalable, and flexible.

Another model that deserves additional analysis is Logistic Regression. This model satisfies many of our goals, has high predictive power, and will allow us to identify the impact of specific attributes on ROI. This is valuable for both us and our end users as it will outline which attributes contribute the most (and least) to a given college’s ROI value. Overall, we believe that our current methodology has led us to a model that both satisfies our outlined goals and generates a relatively high accuracy.

**Evaluation**

The results from the Random Forest models accomplish our goal for a simple model with relatively high accuracy. The accuracy of 69.18% is reasonable for the preliminary evaluation of the model. However, there is room for improvement on the model which we will face in the coming weeks.

We have conducted X-validation on our model on the training set. The result of the validation is displayed in Appendix 7. As explained in previous sections, an accurately quantified cost matrix will be hard to calculate since student’s select college not solely on ROI, but includes other subjective factors. However, we do believe that the cost will be higher for predicting a school to have high or medium ROI when its ROI is actually low because it will incur long-term financial losses for students. Therefore, when developing the model for the data product, our goal is to be as conservative as possible when predicting a high or medium investment for an institution.

Furthermore, given the size of the data set, the data preparation was very costly in terms of time. We encountered tradeoffs in our data preparation and cleaning. For example, we used 2011 data instead of 2013 data because while 2013 data was more recent and relevant, 2011 provided the most accurate and complete data set. For future projects, it would be best to focus on other ways of obtaining data such as using third party data source to complete missing data.

**Recommendations**

The end goal of this project is to provide a dynamic recommender system for high school students to identify colleges with the best ROI based on their interests. Therefore, we will not produce a single recommendation at the end of the project. However, our model has revealed a few interesting misconceptions about ROI in college.

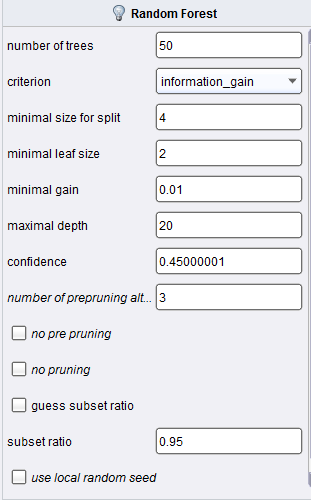
Same factors apply for admission rate. Even though a very high admission rate is generally an indicator of a low return on investment, the situation becomes much more unclear when comparing between schools of high and medium ROI. College with very low admission rate, such as Harvard, only provides medium return on investment while University of Virginia, an institution with a much higher admission rate, provides high return on investment for students.

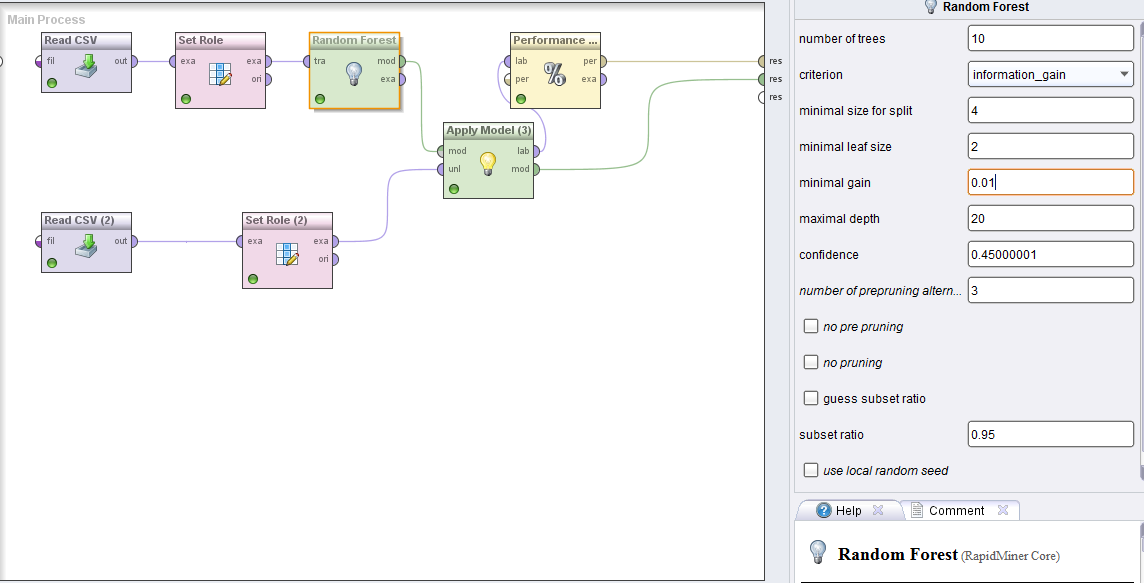
From those observations, we have concluded that tuition has a much bigger impact on ROI than many applicants may realize. The reason the most prestigious institutions have low ROIs is because the prestigious institutions have a huge cost of tuition associated with them which decreases the ROI in relation to other colleges that are less expensive. Interestingly, many community colleges often have a high ROI because of the low tuition cost, even though the median income six years later might not be as high as those graduated from prestigious institutions.

However, this does not mean we will recommend all students go to community colleges simply for the lower tuition--a student’s college preferences must be included in the recommendation for it to be comprehensive.

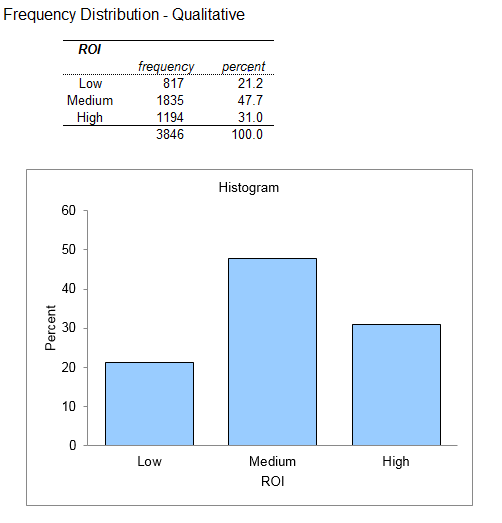
**Appendix 1: Random Forest Parameters**

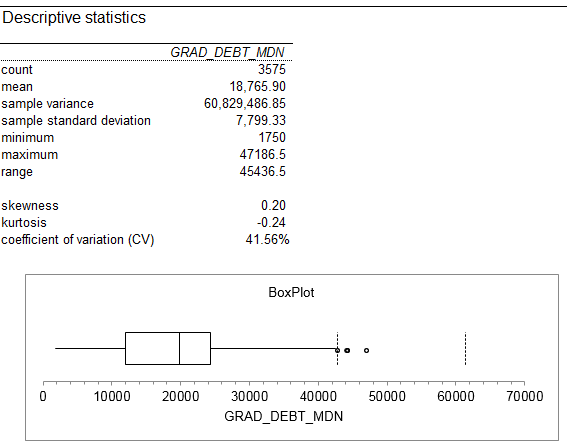
Accuracy: 67.23%

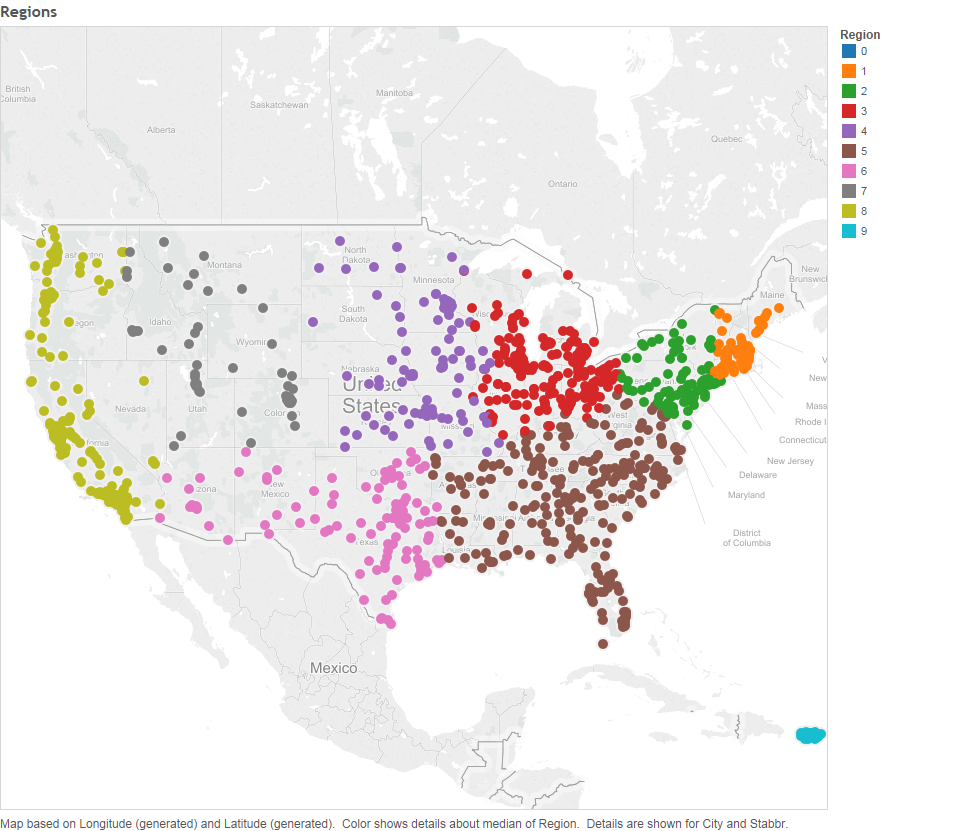
****

**Appendix 2: Random Forest Process**

**Appendix 3: ROI Variable Frequencies**

****

**Appendix 4: Student Debt Descriptive Statistics**

**Appendix 5: Regions in the United States**

**Appendix 6: Summary Statistics**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable name | Mean | Median | Min | Max | Standard Deviation | Missing |
| Admission Rate: The percentage of applicants who are accepted into the 4-year institution | 66% | 68% | 0% | 1900% | 19% | 49% (1867) |
| SAT Math: The colleges’ medians of their SAT math scores | 531 | 520 | 350 | 780 | 70 | 67% (2586) |
| SAT Reading: The colleges’ medians of their SAT reading scores | 524 | 515 | 285 | 750 | 68 | 68% (2609) |
| Graduation Debt: The median of the amount of debt that students graduate with | 18,700 | 19,700 | 1,700 | 47,000 | 7,700 | 7% (271) |
| Undergraduate Students: The number of undergraduate students at the institution | 4,100 | 1,700 | 0 | 250,000 | 7,400 | 0% (0) |
| Average Income: The average familial income of students at the institution | 43,000 | 34,000 | 7,200 | 142,000 | 25,000 | 2% (58) |

**Appendix 7: Performance Vector X-Validation**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **true Medium** | **true High** | **true Low** | **class precision** |
| **pred. Medium** | **1118** | **206** | **337** | **67.31%** |
| **pred. High** | **216** | **734** | **12** | **76.30%** |
| **pred. Low** | **129** | **20** | **305** | **67.18%** |
| **class recall** | **76.42%** | **76.46%** | **46.64%** |  |