Predicting Student Earnings

How the Academic and Non-Academic Qualities of a Student's School Effects Future Income

Prepared by:
Michael McGeehan
October 2017

Microsoft Data Science Capstone Challenge
https://www.datasciencecapstone.org/competitions/2/student-earnings/page/5/

Executive Summary

Leveraging publicly available data from the United States Department of Education, this report seeks to determine a student's future income prospects based on the selection of educational institution and academic program selection. There is a large quantity of data to be used in this analysis; over 17,000 observations and 299 variables. These variables represent information about the educational institutions, existing students, demographics and academic programs offered.

Given the large number of variables, the first step in the analysis was to determine which of those variables had the highest correlation to the income label linked to those variables. This determination was achieved by reviewing summary statistics of each of the variables and the relationship to income. The data was then prepared in charts and graphs to further explore how income is impacted by data contained in the variables. In order to discover how the various variables influenced future income of the prospective students, several regression techniques were tested to determine the optimal predictive ability. The data contains many missing values in the variables which had to be addressed to more accurately predict the future income labels.

After some review, the data revealed information that could provide some practical guidance for prospective students when choosing a higher education institution and an academic major.

The most significant data points based on the analysis include:

School Specific Variables:

Faculty Salary – The salary of the school's staff had the greatest positive influence over the future income of students. On average, Public Colleges paid the faculty the highest salaries but Private Non-Profit colleges paid their faculty the most.

Cost of Tuition – The tuition fees, both in-state and out-of-state, also had a significant positive influence over the students' future income. Cost of private institutions, particularly Private Non-Profit colleges were higher than public schools.

Instructional Expenditures – Unsurprisingly, the Private Non-Profit schools have the highest average instructional expenditure. This expenditure is positively related to future income. However, there are many public schools that spend significantly above the average and thus a few public schools have the highest instructional expenditure per student.

Student Demographic Variables:

"First Generation" Students Whose Parents Went to College – Students that have parents that went to college had the highest correlation to future income for the student. Unfortunately for the student, they have little control over this variable.

First Generation College Students – Students who's parents did not complete high school were the most likely to earn the lowest incomes in the future. There is a strong negative correlation between first generation students at an institution of higher learning and the income they will receive. Even those with parents who completed high school but did not attend college are very likely to earn the lowest future income. Again, the student has no control over this variable.

Academic Variables:

STEM Majors – Students who majored in academic programs that focus on Science, Technology, Engineering and Math had a greater likelihood of a higher future income. Health, Biology, Computer, Business and Mathematics were in the Top 10 of positive correlations.

Liberal Arts Majors – Students matriculating in English, History, Social Sciences and others do not have as high a positive correlation with income as STEM programs but because these majors are found primarily at Non-Profit and Public colleges, graduates will do better than with most programs offered at For Profit schools.

Culinary Certificates – Students that attend school programs for culinary studies have a negative correlation with future income. This is primarily because very few students in these courses attend Non-Profit institutions. Those that attend For-Profit schools attain only modest incomes.

Initial Data Exploration:

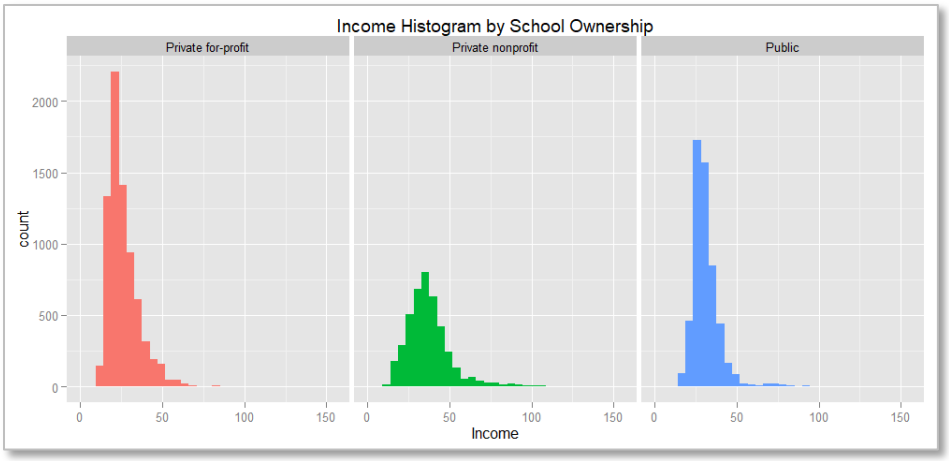
Given that large number of variables that were available for this analysis, it was necessary to determine which ones had the most significant effect on income. This required examining the summary and descriptive statistics of the variables.

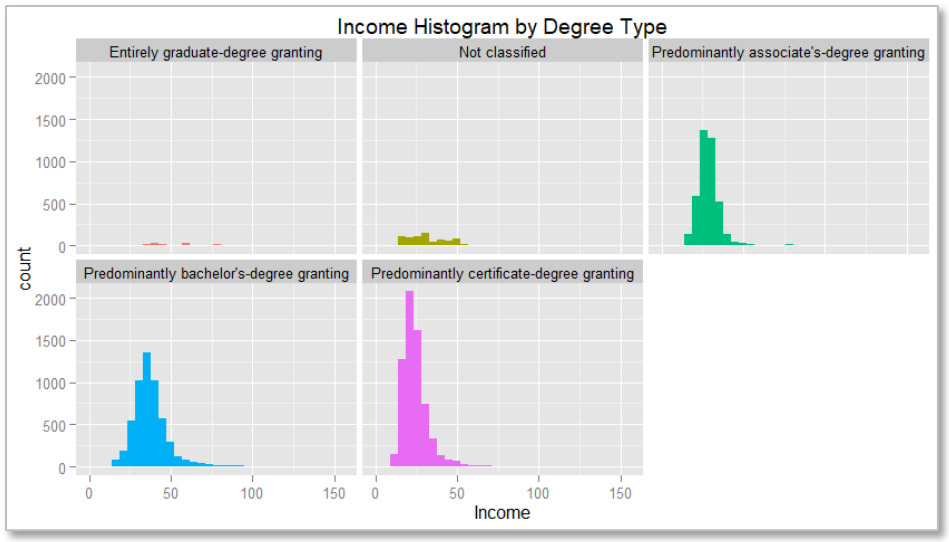
The primary interest of this analysis is the future income based on the selection of school and academic choices of the student. Most of the data points for income congregate around "25" though for Private Non-Profit schools, it's closer to "30".

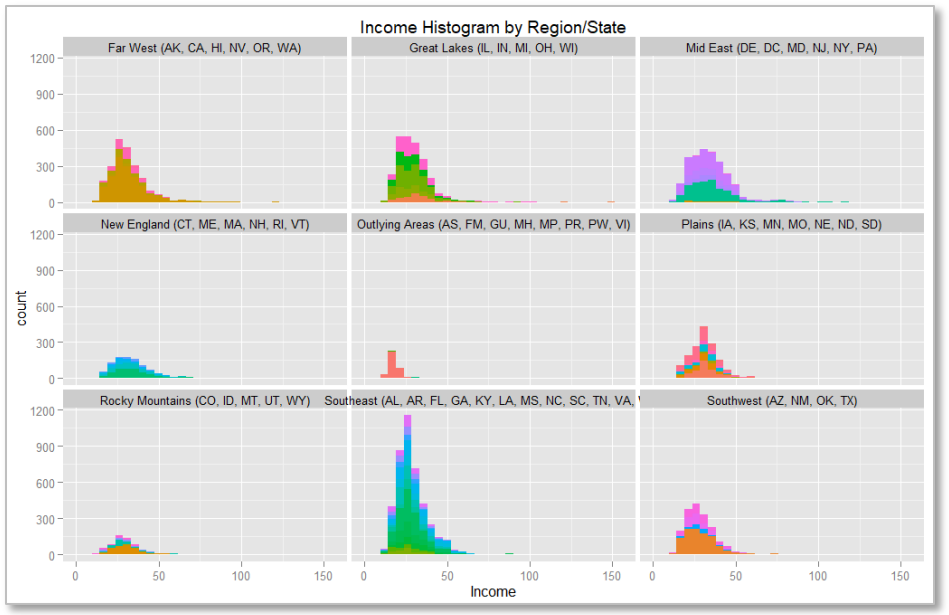
When viewed by Degree Type, income data points cluster in the Bachelors and Associate Degree categories and most frequently, in the Certificate category.

Finally, the income data was derived primarily from the Southeast United States.

These three charts indicate that the best predictions of income will correlate to students attending Public or For-Profit schools in the Southeast getting Associates or Certificate degrees.



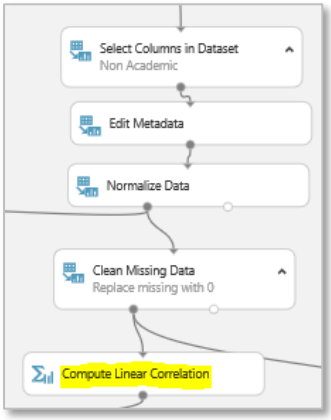




Histograms of Income Values

Non-Categorical Variable Selection:

In order to determine which non-categorical variables should receive the greatest focus, a linear correlation was performed between all the variables and the income label. This was done using the "Compute Linear Correlation" statistical function component in Azure Machine Learning.



To further facilitate the analysis, the non-categorical variables were separated into non-academic and academic categories. The non-academic categories include school specific variables such as School Type, Faculty Salary, Tuition Fees, Admissions Test data, etc. Student specific variables with demographic information such as First Generation at College and Parent's Education attainment level.

A sample of the "Compute Linear Correlation" for the non-categorical Non-Academic variables is shown in the table below:

cost__tuition_out_of_state	school__faculty_salary	school__instructional_expenditure_per_fte	student__demographics_dependent	student__demographics_female_share	student__demographics_first_generation	student__share_firstgeneration	student__share_first generation_parents_highschool	student__share_first generation_parents_somecollege	income
0.01	0.02	0.00	0.01	-0.01	-0.01	-0.01	-0.01	0.01	0.00
0.36	0.33	0.14	0.09	-0.04	-0.28	-0.28	-0.19	0.29	0.21
0.55	0.39	0.11	0.61	-0.19	-0.62	-0.62	-0.50	0.62	0.36
0.33	0.32	0.13	0.09	-0.05	-0.25	-0.25	-0.18	0.26	0.21
0.34	0.35	0.14	0.08	-0.06	-0.27	-0.27	-0.20	0.28	0.23
0.33	0.33	0.13	0.08	-0.07	-0.26	-0.26	-0.20	0.28	0.21
0.37	0.34	0.14	0.11	-0.07	-0.31	-0.31	-0.23	0.33	0.23
0.37	0.34	0.14	0.10	-0.06	-0.30	-0.30	-0.20	0.31	0.22
0.34	0.35	0.14	0.08	-0.07	-0.27	-0.27	-0.20	0.28	0.23
0.92	0.21	0.13	0.24	-0.04	-0.43	-0.43	-0.25	0.44	0.27
1	0.39	0.15	0.37	-0.07	-0.52	-0.52	-0.36	0.53	0.32
0.39	1	0.19	0.38	-0.12	-0.41	-0.41	-0.42	0.43	0.41
0.15	0.19	1	0.10	-0.04	-0.15	-0.15	-0.11	0.16	0.23
0.37	0.38	0.10	1	-0.34	-0.67	-0.67	-0.53	0.66	0.20
-0.07	-0.12	-0.04	-0.34	1	0.25	0.25	0.20	-0.25	-0.23
-0.52	-0.41	-0.15	-0.67	0.25	1	1.00	0.77	-0.96	-0.46
-0.52	-0.41	-0.15	-0.67	0.25	1.00	1	0.77	-0.96	-0.46
-0.36	-0.42	-0.11	-0.53	0.20	0.77	0.77	1	-0.73	-0.42
0.53	0.43	0.16	0.66	-0.25	-0.96	-0.96	-0.73	1	0.44

Partial Output of Non-Academic variables from Compute Linear Correlation Component in Azure Machine Learning

A sample of the "Compute Linear Correlation" for the non-categorical Academic variables is shown in the table below:

academics__program_percentage_biological	academics__program_percentage_business_marketing	academics__program_percentage_engineering	academics__program_percentage_history	academics__program_percentage_mathematics	academics__program_percentage_personal_culinary	academics__program_percentage_psychology	academics__program_percentage_social_science	income
1.00	0.08	0.13	0.48	0.44	-0.16	0.27	0.43	0.24
0.08	1.00	0.02	0.11	0.09	-0.29	0.13	0.07	0.24
0.13	0.02	1.00	0.08	0.27	-0.08	0.04	0.11	0.25
0.48	0.11	0.08	1.00	0.58	-0.18	0.38	0.64	0.26
0.44	0.09	0.27	0.58	1.00	-0.16	0.30	0.52	0.27
-0.16	-0.29	-0.08	-0.18	-0.16	1.00	-0.16	-0.16	-0.38
0.27	0.13	0.04	0.38	0.30	-0.16	1.00	0.32	0.21
0.43	0.07	0.11	0.64	0.52	-0.16	0.32	1.00	0.28
0.24	0.24	0.25	0.26	0.27	-0.38	0.21	0.28	1.00

Output of Academic variables from Compute Linear Correlation Component in Azure Machine Learning

The non-categorical variables where selected based on the absolute value of the correlation coefficient to the Income label.

Non-Academic Variable Name	Correlatior
student__demographics_first_generation	-0.455
student__share_firstgeneration	-0.455
student__share_firstgeneration_parents_somecollege	0.432
student__share_firstgeneration_parents_highschool	-0.417
school__faculty_salary	0.407
cost__tuition_out_of_state	0.319
cost__tuition_in_state	0.269
school__instructional_expenditure_per_fte	0.249
admissions__sat_scores_25th_percentile_math	0.226
admissions__sat_scores_midpoint_math	0.224
admissions__sat_scores_average_by_ope_id	0.223
admissions__sat_scores_average_overall	0.216
admissions__sat_scores_75th_percentile_math	0.213
admissions__act_scores_25th_percentile_cumulative	0.210
admissions__act_scores_midpoint_cumulative	0.205
admissions__act_scores_25th_percentile_math	0.204
admissions__act_scores_midpoint_math	0.204
student__demographics_dependent	0.201

Non-Categorical Academic Variable Names	Correlation
academics__program_percentage_personal_culinary	-0.381
academics__program_percentage_social_science	0.279
academics__program_percentage_mathematics	0.266
academics__program_percentage_history	0.264
academics__program_percentage_engineering	0.253
academics__program_percentage_biological	0.242
academics__program_percentage_business_marketing	0.240
academics__program_percentage_psychology	0.207

Non-Categorical variables selected for the Income analysis

Categorical Variable Selection:

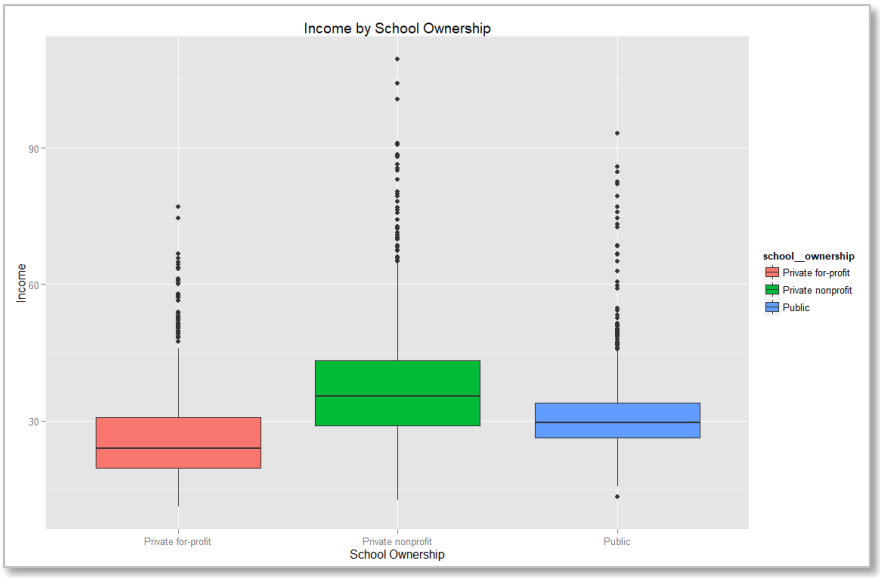
The Categorical Academic variables have only four values; 0, 1, 2 and Null. Before turning these variables into Categorical variables using the Edit Metadata functionality in Azure Machine Learning, a numerical correlation was made with the Income label in order to determine the optimal variables to include in the analysis. The list of those variables is below:

Academic Variable Names	Correlation
academics__program_certificate_lt_2_yr_personal_culinary	-0.370
academics__program_bachelors_health	0.341
academics__program_bachelors_computer	0.336
academics__program_bachelors_psychology	0.332
academics__program_bachelors_business_marketing	0.322
academics__program_bachelors_social_science	0.315
academics__program_bachelors_biological	0.315
academics__program_bachelors_multidiscipline	0.309
academics__program_bachelors_mathematics	0.309
academics__program_certificate_lt_1_yr_personal_culinary	-0.307
academics__program_bachelors_physical_science	0.307
academics__program_bachelors_english	0.307
academics__program_bachelors_history	0.296
academics__program_bachelors_philosophy_religious	0.292
academics__program_bachelors_language	0.292
academics__program_bachelors_engineering	0.278
academics__program_bachelors_humanities	0.278
academics__program_bachelors_communication	0.277
academics__program_bachelors_education	0.277
academics__program_bachelors_visual_performing	0.267
academics__program_bachelors_ethnic_cultural_gender	0.258
academics__program_bachelors_public_administration_social_service	0.255
academics__program_bachelors_resources	0.238
academics__program_certificate_lt_2_yr_health	-0.201

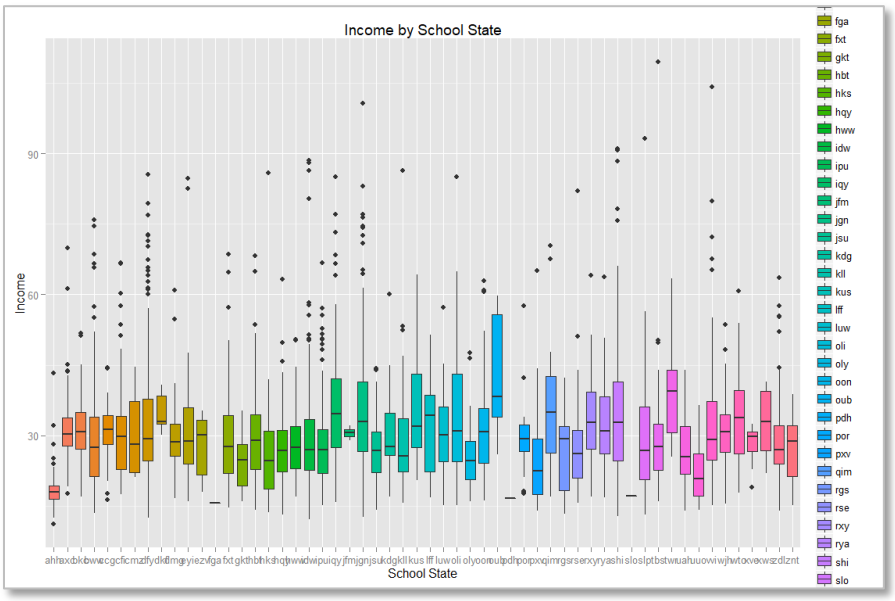
Categorical Academic Variables Selected based on Correlation to Income

Categorical Non-Academic variables were chosen based on individual analysis of the variables using boxplots.

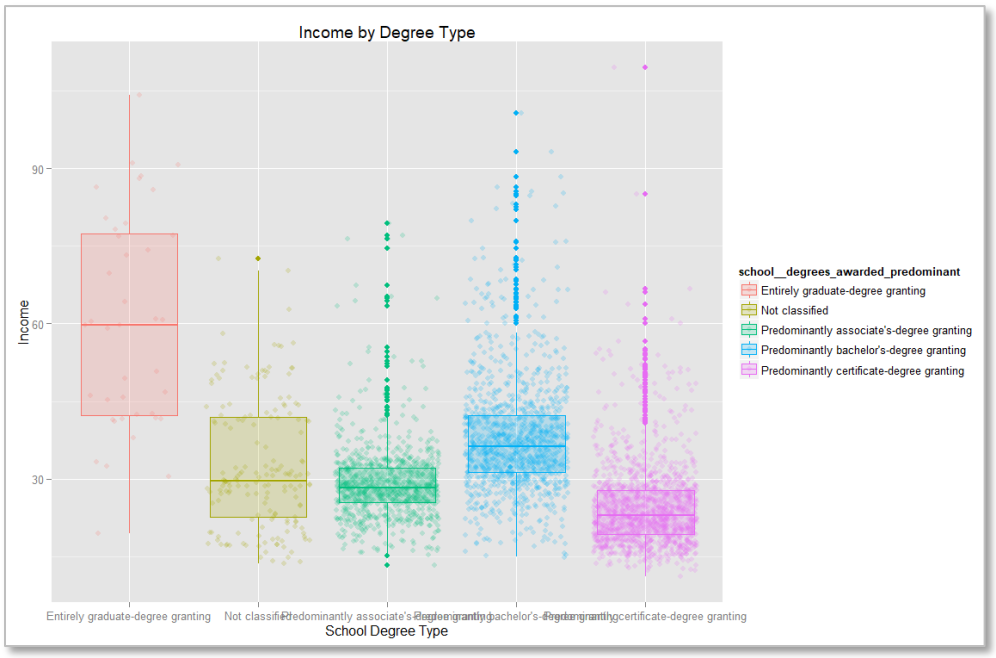
School Ownership: A boxplot comparison shows that students attending Private Non-Profit schools had the highest mean income and the highest income potential. Students from Public colleges had a slightly lower mean and range but still a fairly high potential income. Private For-Profit schools had the lowest mean and income potential.



School State: The state where the educational institution also has an effect on income. Most states overlap with respect to range but there is one standout state, "oub", that has the highest impact on income.

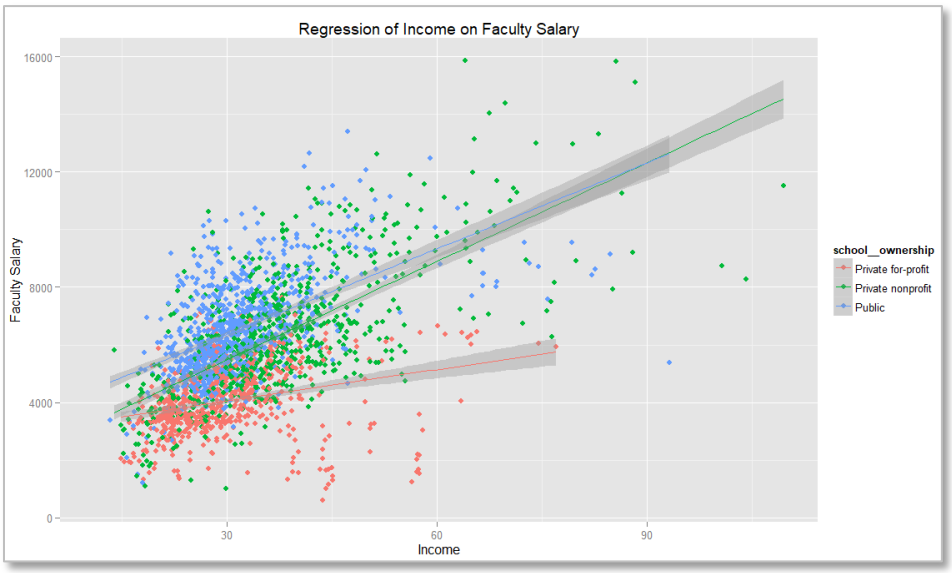


Degree Type: The type of degree predominately issued by the academic institution also correlates to income. Though colleges that focus primarily on graduate degrees have the highest average income, very few students attend those types of schools as shown in the chart below.



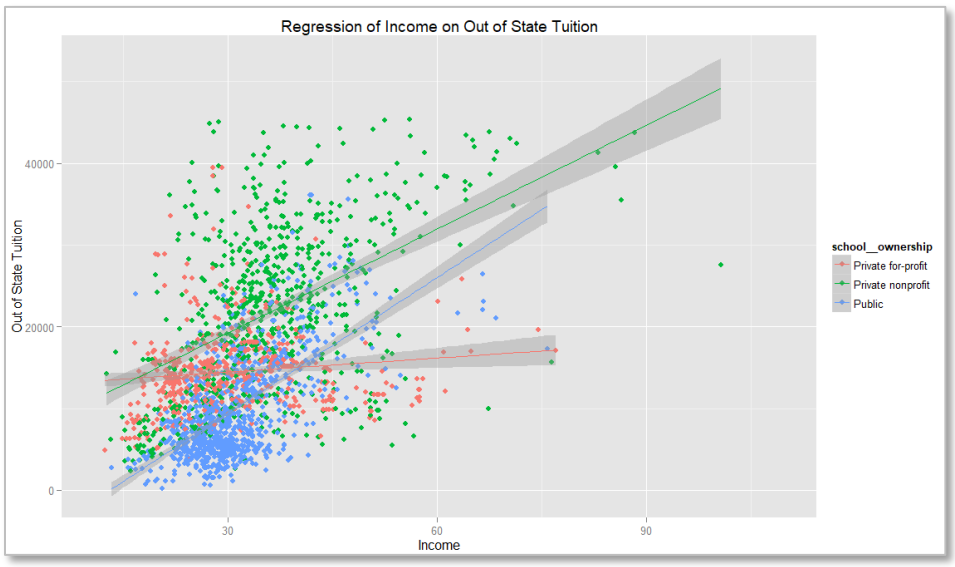
Regressions and Relationships with Non-Categorical, Non-Academic Variables

Faculty Salary: There is a high, positive correlation between the earnings of school faculty and the income of graduates. This relationship applies for all school no matter the ownership.



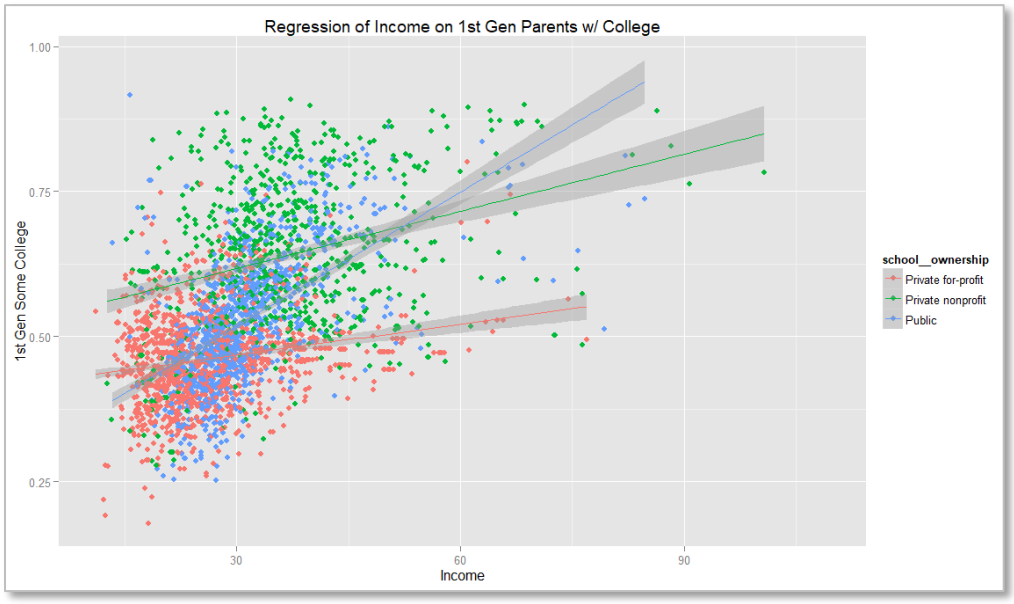
Scatter Plot of Income versus Faculty Salary shown by School Ownership

Tuition Costs: There's also a high, positive correlation between the cost of the academic institution and income. The relationship with income is most pronounced with Public Schools where both degree types, Associate's Degrees and Bachelor's Degrees are offered with high frequency. Students attending Private For-Profit schools do not benefit as much from paying higher tuitions.

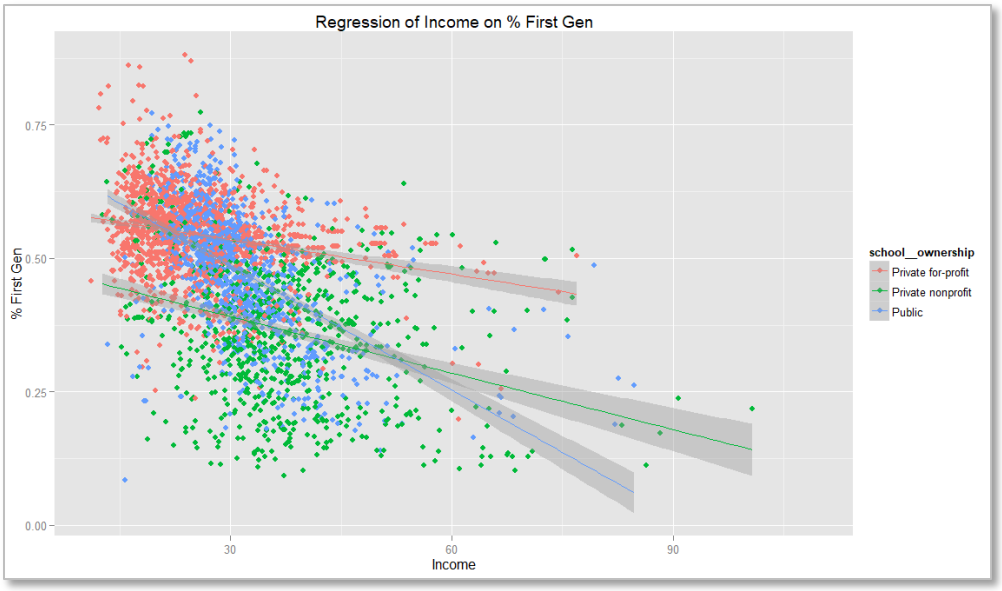


Scatter Plot of Income versus School Tuition shown by School Ownership

Students with Parents who Attended College: Another high correlation with income appears when looking at students that have parents that have attended some college. We can conclude that such students are better prepared for the academic rigors compared to first generation college students. The contrast can be seen in the two charts below. While there is a very positive correlation between income and students with parents that attended college, there is a very negative correlation between income and students who are the first generation to attend college. Unfortunately, unlike school selection and academic program choice, the student is not able to influence this statistic.



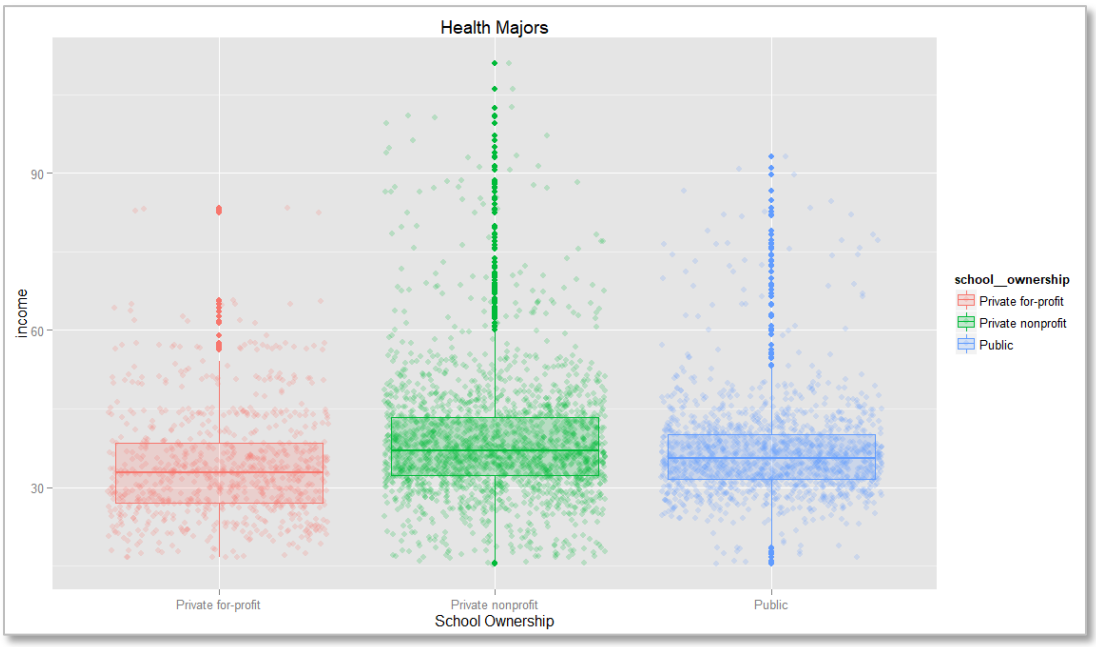
Scatter Plot of Income versus Students with Parents that Attended College shown by School Ownership



Scatter Plot of Income versus First Generation Student Attending College shown by School Ownership

Regressions and Relationships with Non-Categorical, Academic Variables

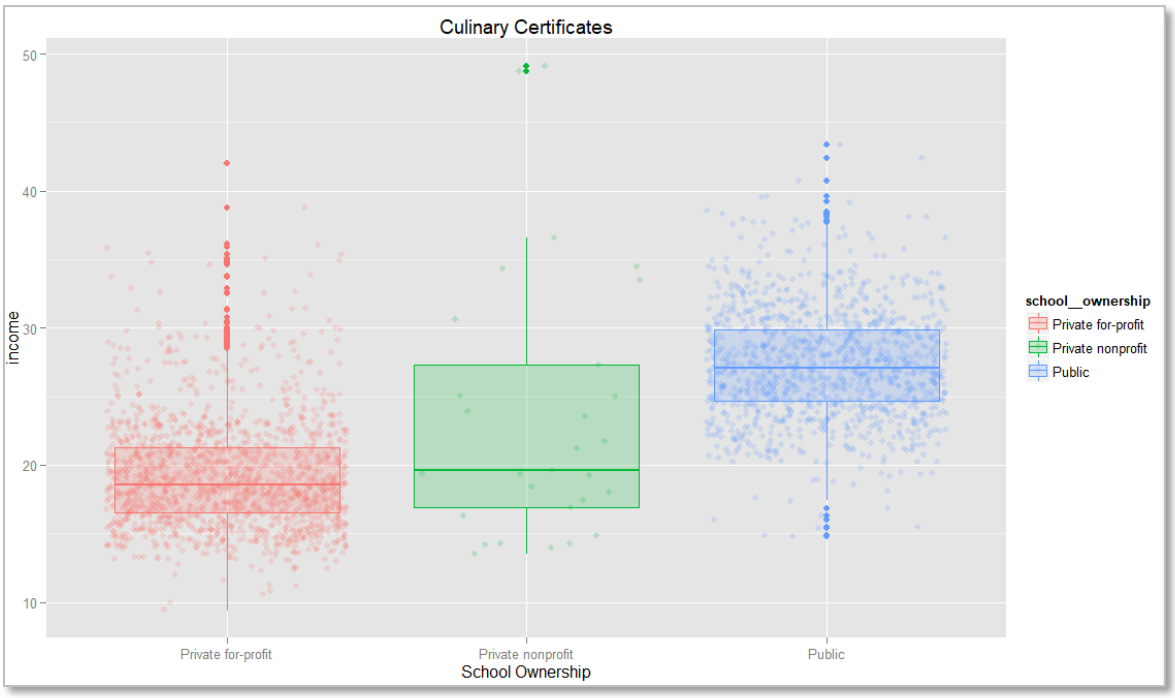
Health Majors: Students that majored in health related fields had the highest positive correlation with income. This finding was consistent across all school ownership types. The highest paid health majors achieved at least a Bachelor's degree but students who attained an Associate's degree or certificate also reflected higher income potential.



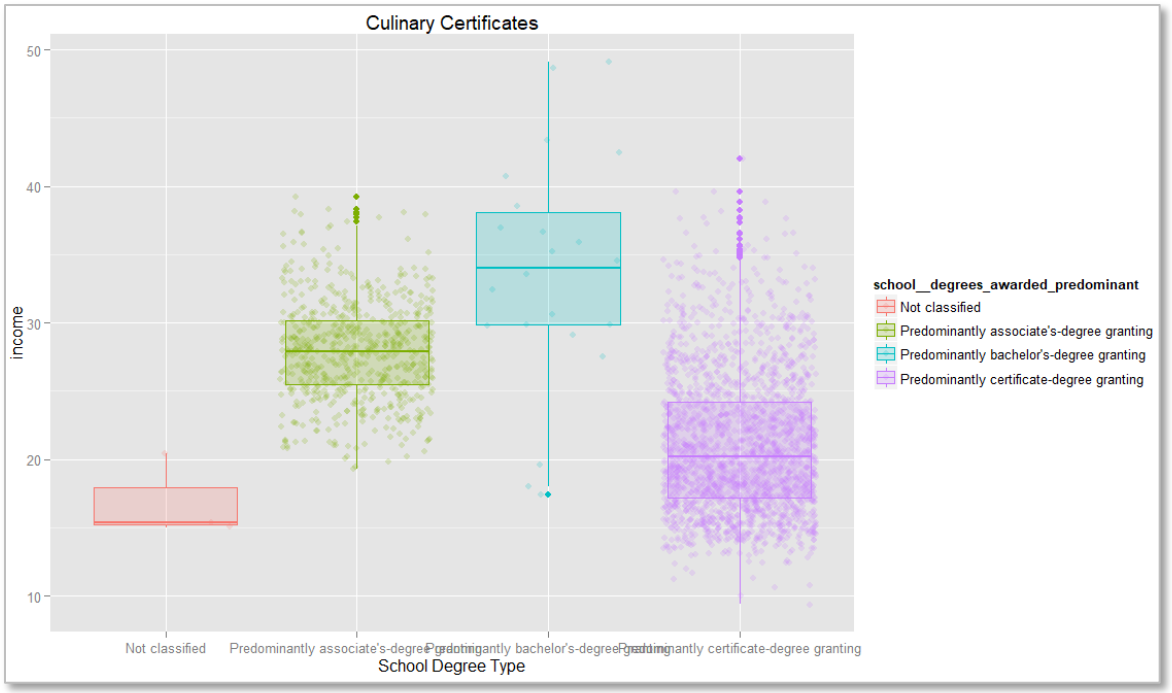
Income of Health Majors shown by School Ownership

Income of Health Majors shown by School Degree Type

Culinary Majors: On the other end of the spectrum, students that attained certificates in culinary programs clustered at the lowest income levels. Those who attended public schools where an Associate's Degree was offered had significantly higher income than those that attended Private For-Profit schools earning a certificate.



Income of Culinary Majors shown by School Ownership



Income of Culinary Majors shown by School Degree Type

Regression Analysis:

Based on the understanding of the variables and their relationship to income, an Azure Machine Learning model was created to predict income of a test dataset based on data in the training dataset.

There were several significant choices to make when creating the model:

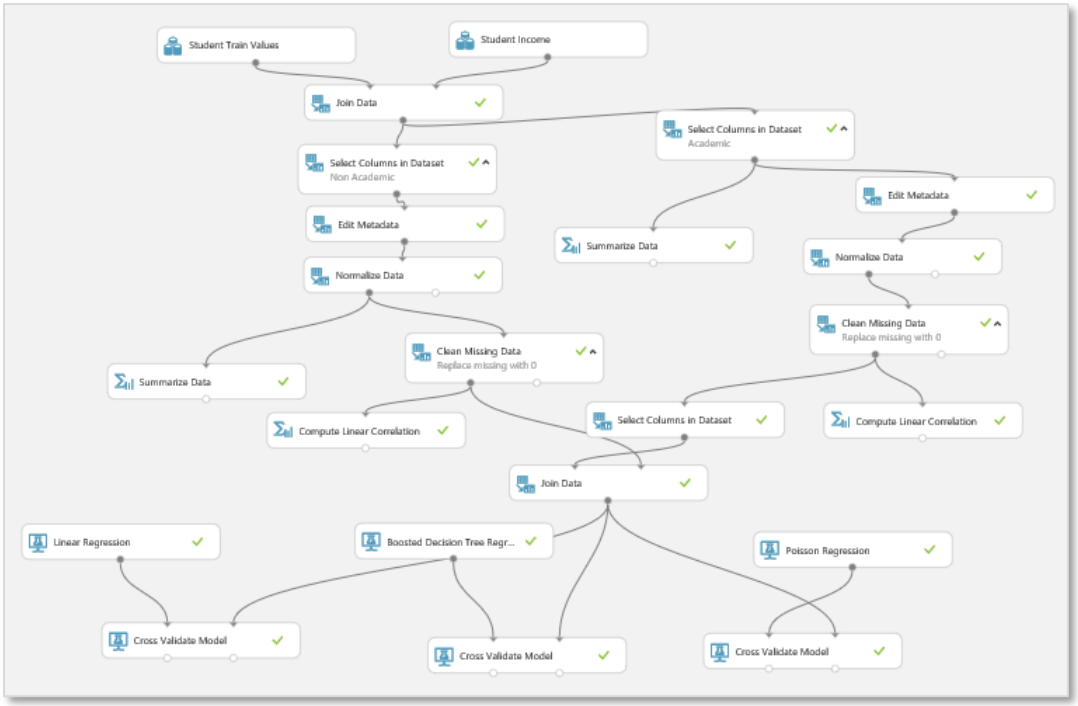
- Which Regression Model to use
- How to address missing values in the data for the selected variables
- How to handle outlying values in the data for the selected variables

Regression Model Selection:

Using the training data, three different regression models were tested.

- Linear Regression
- Boosted Decision Tree Regression
- Poisson Regression

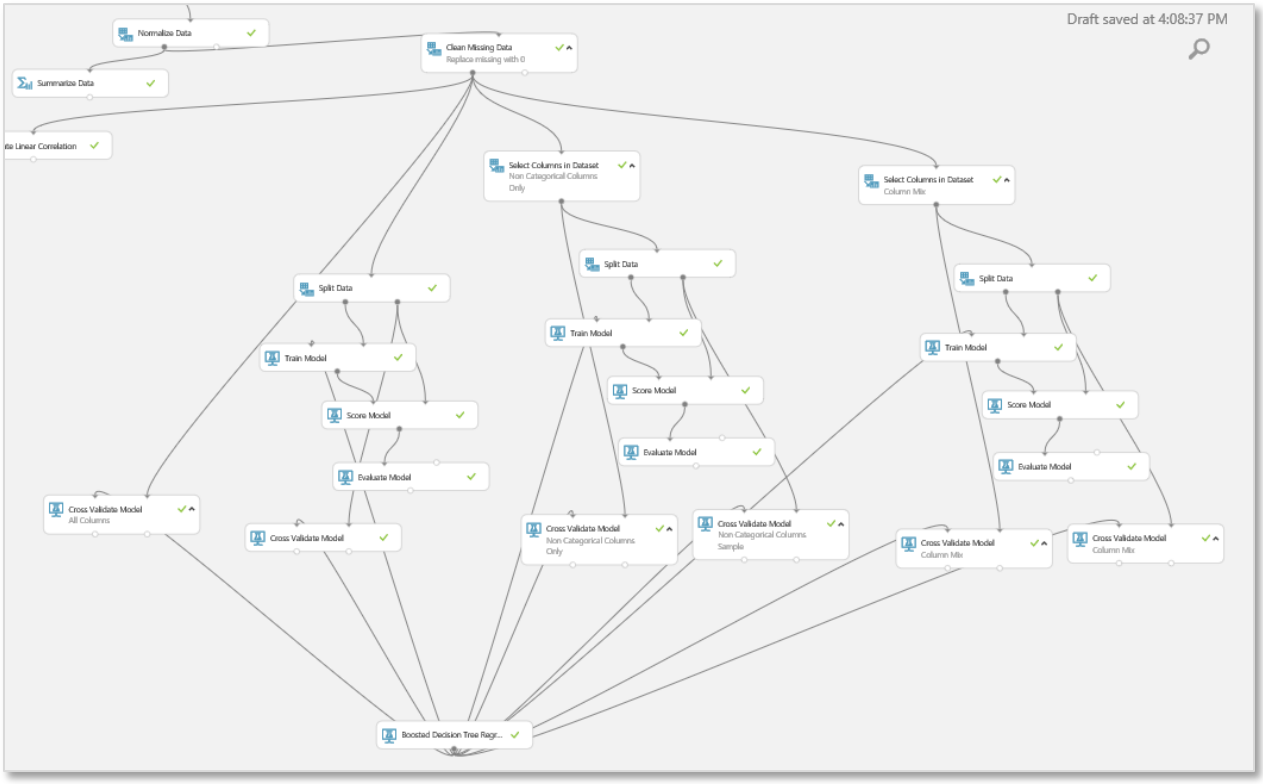
The results of each regression were tested under several different conditions. At times, missing data was substituted with a "0" value. Other times, the missing data was substituted with mean values. Under all circumstances, it was found that the Boosted Decision Tree Regression had the lowest Root Mean Squared Error values. For this reason, the Boosted Decision Tree was selected as the regression model for use in the analysis.



Azure Machine Learning Experiment Testing Various Regression Methods

Column Selection and Missing Data:

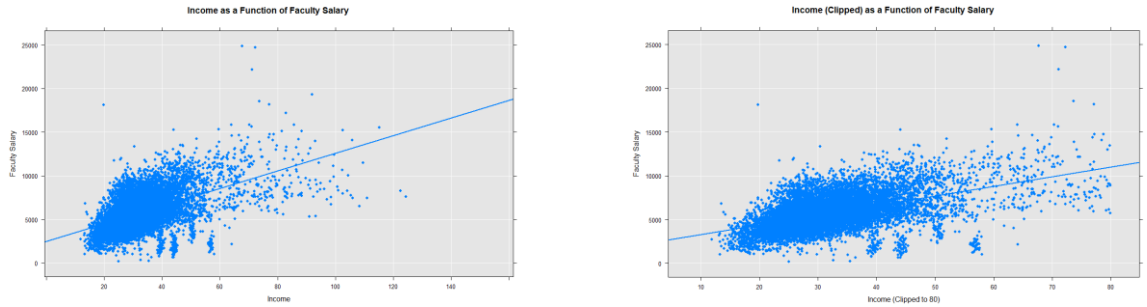
Once the Boosted Decision Tree Regression method was selected, various attempts were made to optimize the column selection and infilling of missing data. Different methods of filling missing data were tried for Academic and Non-Academic variables; Categorical and Non-Categorical. After many different parameters were tried, it was decided that filling in the missing variable data with "0" led to the optimal regression fitting of the test data.



Azure Machine Learning Experiment Optimizing the Boosted Decision Tree Regression

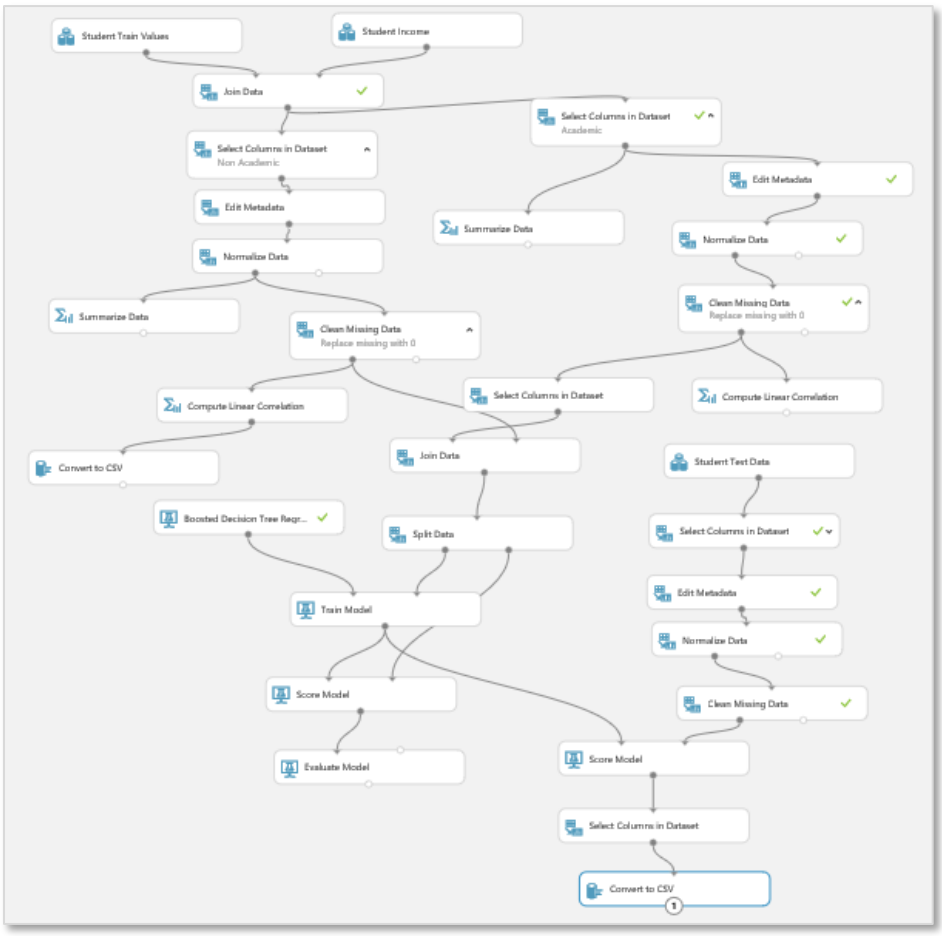
Outlying Values:

There were many datapoints on the high end of the income label that had significant leverage over the regression line plotted against several other variables. For example:



Change in Slope of Linear Regression Line when Eliminating Income Values Higher than 80

Several Azure Machine Learning experiments were created to test clipping values from income and various other variables. Through trial and error, it was determined that none of the outliers should be clipped in order to achieve the lowest RSME score when running the regression analysis against the test data.



Azure Machine Learning Model that Produced the Lowest RSME Score.

Link to Published Model:

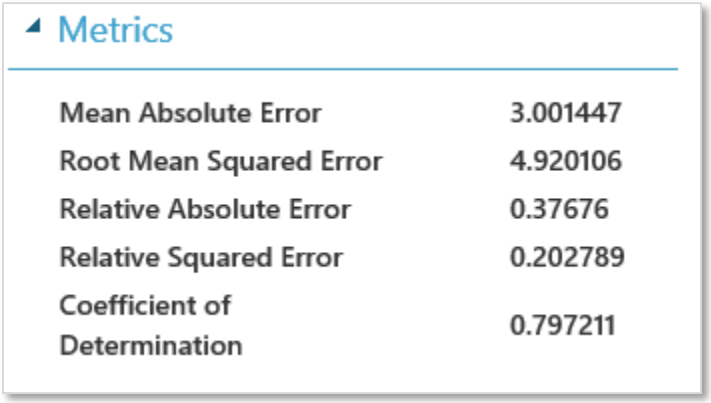
[DAT102x: Predict Student Earnings – Oct 2017 - McGeehan]

Conclusion:

This analysis shows that a student's income can be confidently predicted based on a combination of student demographics, school selection and academic programs. Schools with higher tuition and highly paid faculty will likely lead to a higher income. However, if the student selects a STEM major or a Liberal Arts major at a Private Non-Profit college, they will also benefit from a higher income. Students who represent the first generation to attend a higher education

institution will have a more difficult time achieving a higher income compared to other students whose parents attended college.

The metrics of the final regression experiment are shown below. The RMSE score for the data split from the training data was 4.92. When scored against the test data, the RMSE score was 5.5859.



Further analysis on the replacement of missing values in the various variables may improve the RMSE score.