Analysis of Student Earnings

Nils Rohe, October 2017

Executive Summary

This report presents an analysis of data that was made public by the United States Department of Education. The data contains a variety of variables that might affect the future income of students.

Goal of the analysis is the prediction of student's earnings whereas the target variable is 'income' and reflects the earnings some years after the students have enrolled in US institutions of higher education.

The test-dataset contains 17,107 observations and 298 features including the income. The explanatory variables can be split into the following categories: 'Academics', 'Admission', 'Completion', 'Cost', 'Year', 'School', 'Student'.

As a first step, the data was explored thoroughly by visualizing it and calculating various summary and descriptive statistics. Potential relationships with income were identified.

After imputation of missing values, a gradient boosting regression was applied to predict future earnings which scored an R^2 value of 0.901605287421 and an RMSE of 3.53211744243.

From the data exploration and the regression analysis, it can be concluded that two of the categories are primarily important: 'School' and 'Student'. More precisely, the following features are significantly important when predicting future income:

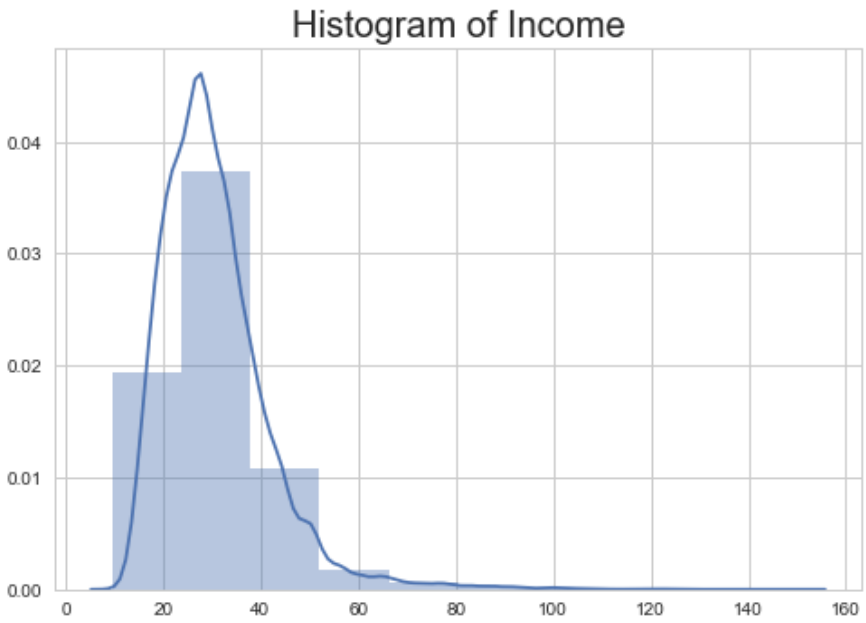
- In the category 'school', monetary aspects are significantly important. The key features are the instructional expenditures per FTE student, the net tuition revenue per FTE student, the average faculty salary and the proportion of faculty staff that is full-time. Hence, the funding of the institution plays a major role and fosters a higher future income.
- In the category 'student', significant features are the age (demographics age entry and share 25 older), the female share, the size, the demographics married, measures of dependence (share independent students and demographics dependent) and also the education of the parents (share first generation, share first generation parents some college, demographics first generation, share first generation parents high school).
  Consequently, it can be concluded that younger, unmarried men that are financially independent and whose parents had a college education themselves tend to have a higher income than other students.

Data Exploration

Income

The variable to predict is the income of students a couple of years after they have enrolled to a US institution of higher education. The distribution and the summary statistics of the variable income in the training dataset are as follows:

Summary Statistics for Income	
Count	17,107
Mean	30.592003
Median	28.70
Std	11.302597
Min	9.40
25%	23.00
50%	28.70
75%	35.60
Max	151.50



As visible from the statistics, mean and median deviate slightly and the standard deviation of 11.3 suggests that there is quite some variance in income. Furthermore, it is apparent from the statistics and the right-skewness of the distribution that there is a small group of people with very high income but a majority with lower incomes.

In the following, not all 297 explanatory variables will be discussed. Rather, the most interesting features will be presented.

Category ‘academics’

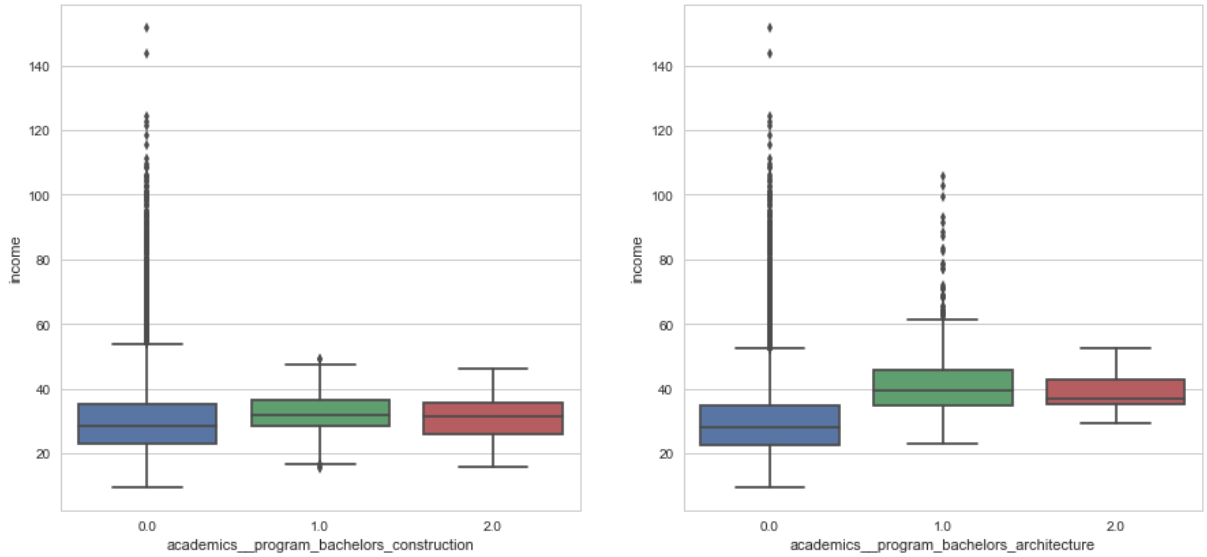
The category ‘academics’ consists of 228 features which indicate which programs each institution offers and the percentage of degrees awarded in each program topic.

Hence, the dataset indicates for each institution whether they offer either an associate degree, a bachelor’s degree, a certificate of less than one academic year, a certificate of at least one but less than two academic years, a certificate of more than two but less than four academic years, or any combination of the above, in one or multiple of the following 38 topics:

Agriculture	Architecture	Biological	Business marketing	Communication
Communications technology	Computer	Construction	Education	Engineering
Engineering technology	English	Ethnic cultural gender	Family consumer science	Health
History	Humanities	Language	Legal	Library
Mathematics	Mechanic repair technology	Military	Multidiscipline	Parks recreation fitness
Personal culinary	Philosophy religious	Physical science	Precision production	Psychology
Public administration social service	Resources	Science technology	Security law enforcement	Social science
Theology religious vocation	Transportation	Visual performing		

In addition and as already mentioned above, further 38 features indicate the percentages of degrees awarded per topic.

Although the category has the highest number of features in the dataset it does not seem to necessarily add too much value as it presumably replicates information that is already contained in other features in the other categories. Features in the school category already determine whether the type of the degree influences income. Hence, only the differentiation between topics remains as unique information in this category.



As an example, the two boxplots above investigate the topics construction and architecture. If not offered (0), both topics yield a very similar income. For both topics, an offered bachelor's degree (1) increases the income but much more for architecture. It is also noticeable that architecture produces more outliers with very high income. A program offered through an exclusively distance-education program (2) does not increase the average income anymore.

Due to the high number of features and their expected limited impact not more than the above example will be discussed in this report. However, it is concluded that the topic of studies has a certain impact.

Category ‘admission’ & category ‘completion’

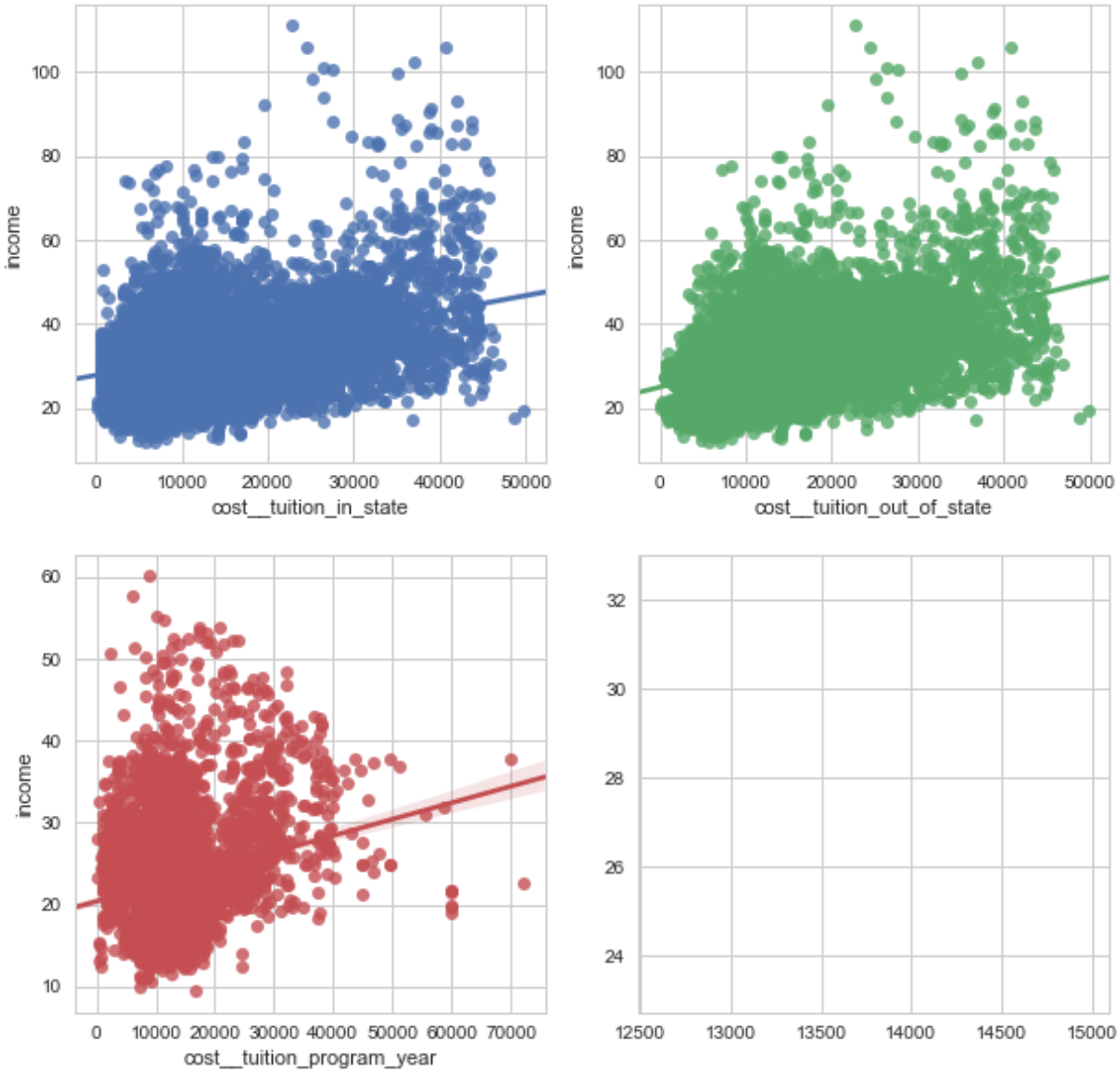
The plots above for selected features of the two categories partially show a clear relationship to income. It appears that especially the SAT and ACT sores are important for future income. The

dataset contains various features concerning these scores and all their plots are in accordance with the above two (plot one and three in the first row).

Despite these promising plots, the two categories are deemed to be of marginal importance as they contain a lot of missing values. The following table shows that almost all features have more than 50% of their values missing (as discussed already the dataset contains 17,107 observations):

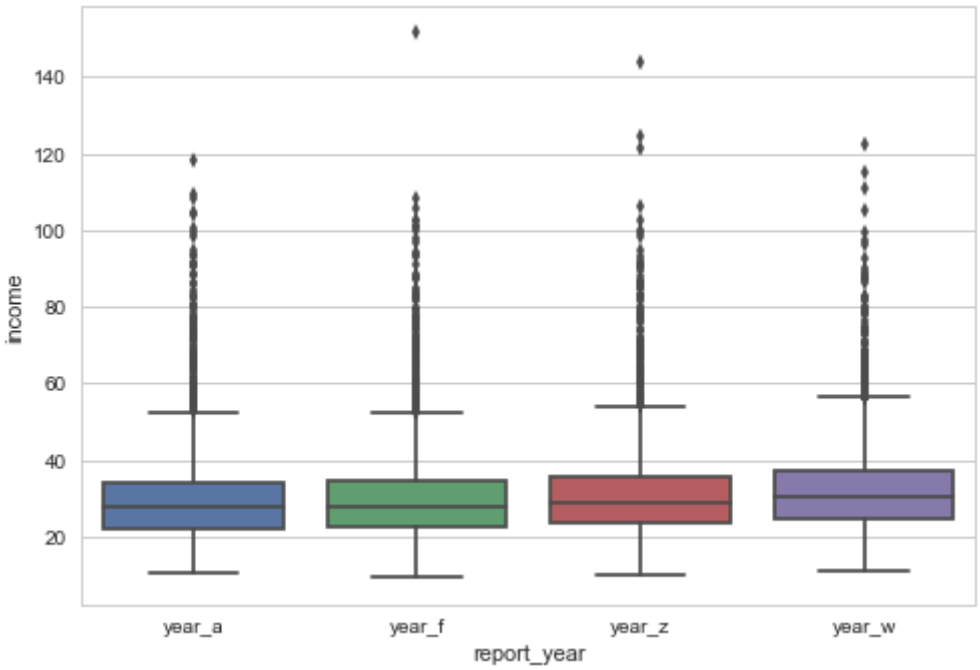
Feature	# of NaNs
admissions__act_scores_25th_percentile_cumulative	13,412
admissions__act_scores_25th_percentile_english	13,969
admissions__act_scores_25th_percentile_math	13,973
admissions__act_scores_25th_percentile_writing	16,676
admissions__act_scores_75th_percentile_cumulative	13,412
admissions__act_scores_75th_percentile_english	13,969
admissions__act_scores_75th_percentile_math	13,973
admissions__act_scores_75th_percentile_writing	16,677
admissions__act_scores_midpoint_cumulative	13,412
admissions__act_scores_midpoint_english	13,969
admissions__act_scores_midpoint_math	13,973
admissions__act_scores_midpoint_writing	16,677
admissions__admission_rate_by_ope_id	9,623
admissions__admission_rate_overall	10,428
admissions__sat_scores_25th_percentile_critical_reading	13,523
admissions__sat_scores_25th_percentile_math	13,477
admissions__sat_scores_25th_percentile_writing	15,120
admissions__sat_scores_75th_percentile_critical_reading	13,523
admissions__sat_scores_75th_percentile_math	13,477
admissions__sat_scores_75th_percentile_writing	15,120
admissions__sat_scores_average_by_ope_id	12,768
admissions__sat_scores_average_overall	13,082
admissions__sat_scores_midpoint_critical_reading	13,523
admissions__sat_scores_midpoint_math	13,477
admissions__sat_scores_midpoint_writing	15,120
completion__completion_cohort_4yr_100nt	11,489
completion__completion_cohort_less_than_4yr_100nt	10,467
completion__completion_rate_4yr_100nt	11,492
completion__completion_rate_less_than_4yr_100nt	10,468
completion__transfer_rate_4yr_full_time	10,910
completion__transfer_rate_cohort_4yr_full_time	10,910
completion__transfer_rate_cohort_less_than_4yr_full_time	8,312
completion__transfer_rate_less_than_4yr_full_time	8,312

Category ‘cost’



The category ‘cost’ contains the three features shown above. ‘cost__tuition_in_state’ and ‘cost__tuition_out_of_state’ almost show identical plots and a positive relationship with income. ‘cost__tuition_program_year’ on the contrary does not show a clear relationship.

Category ‘year’

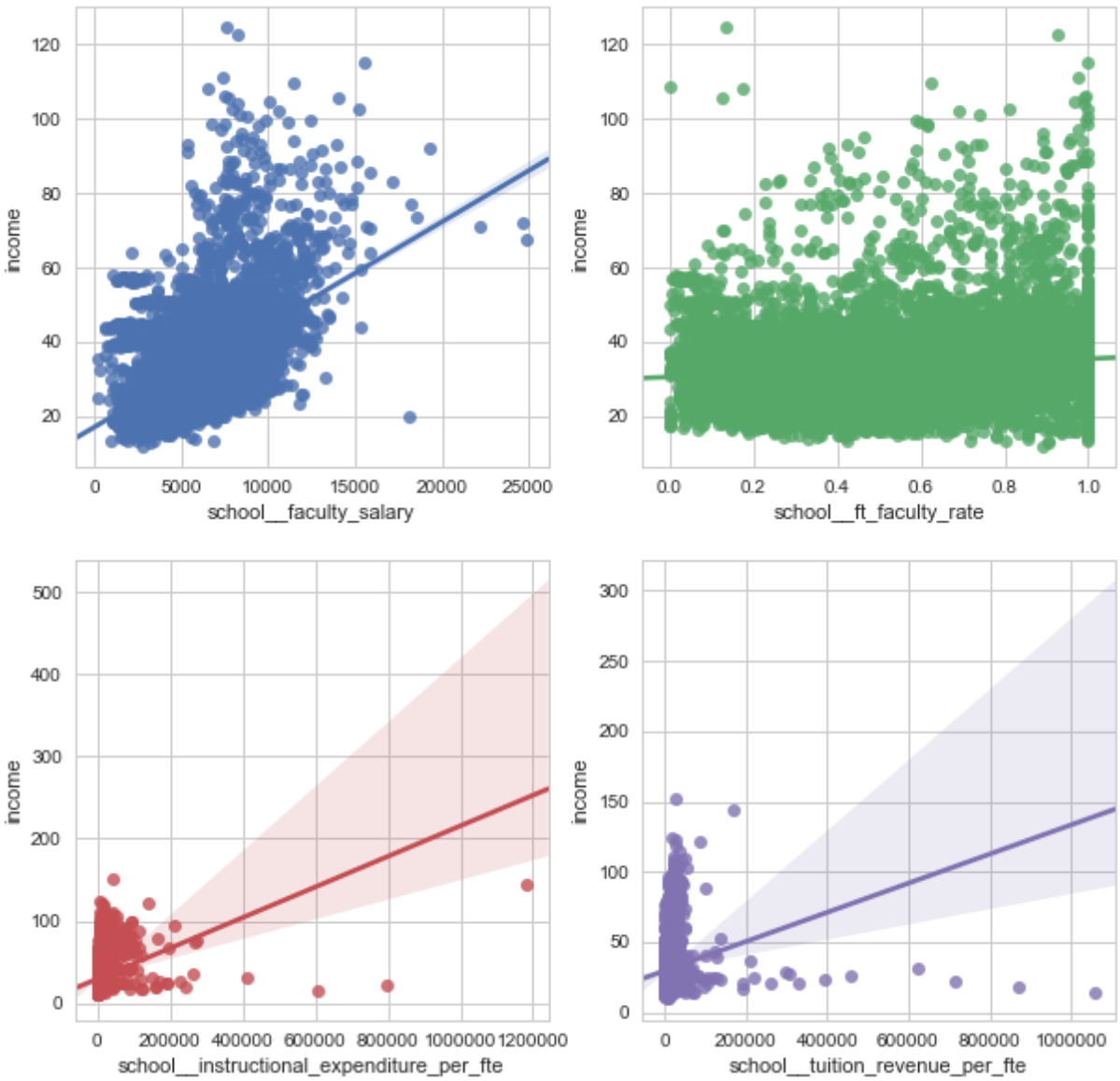


	median	mean	count	std	min	max
year_a	27.8	29.499083	4470	10.928584	10.6	118.4
year_f	27.9	29.892562	4423	11.305736	9.4	151.5
year_w	30.4	32.196949	4032	11.308555	10.8	122.5
year_z	28.75	30.952559	4182	11.494096	10	143.6

The category ‘year’ only consists of the feature ‘report_year’. We can observe from the boxplot and the statistics that the years a and f were very similar and year z just a bit deviated. Year w appears to be the best year in terms of future income with a higher mean and median compared to the other years.

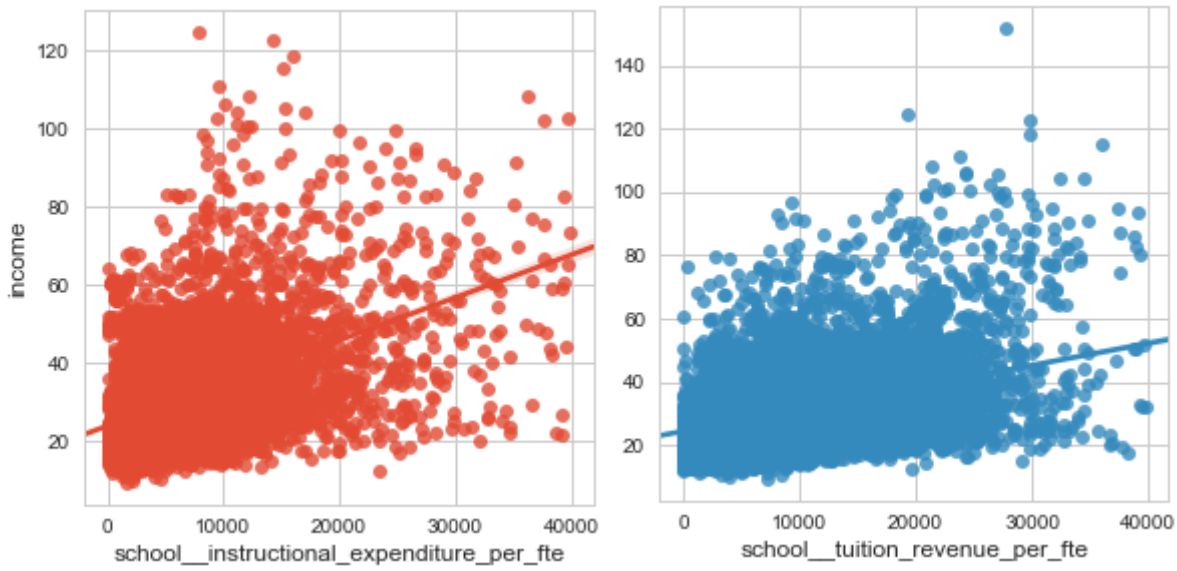
The expectation is that the variable ‘report_year’ only has a small impact on future income.

Category ‘school’

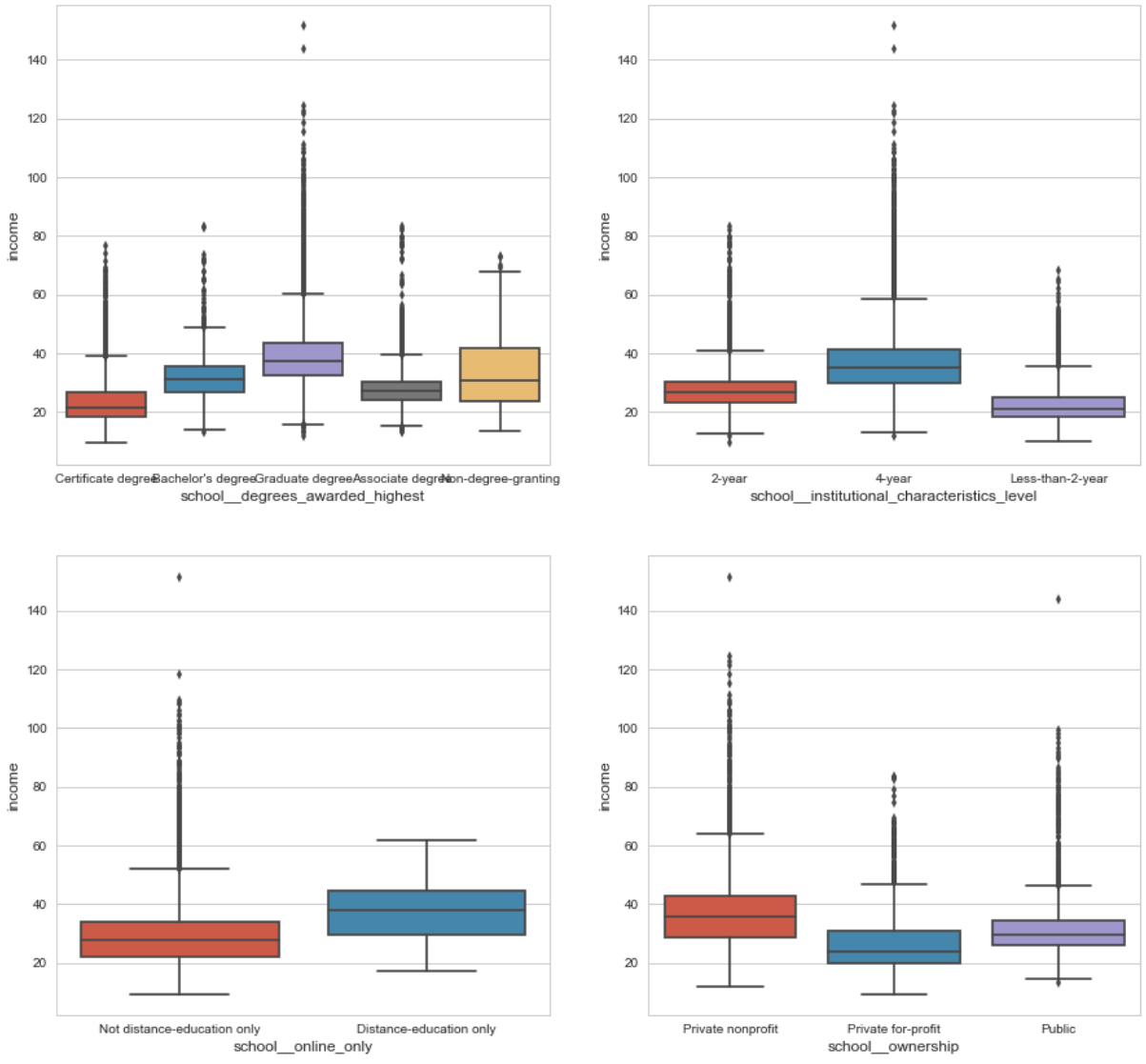


From the numerical features above in the category ‘school’ only ‘school__faculty_salary’ shows a clear positive relationship with income. The faculty rate seems to be unimportant; some extreme outliers in the data distort the other two plots.

However, if we focus on the smaller values until 40,000 for the latter two we can see that there is indeed a relationship with income:



Furthermore, selected categorical features are as follows:



Not surprisingly, we can see in the first boxplot that a higher degree also yields a higher income. This is also supported by the second boxplot showing that a longer education yields a higher future income.

Interestingly, distance-education shows a significant difference to not distance-education. However, the group of distance-education only consists of very few observations:

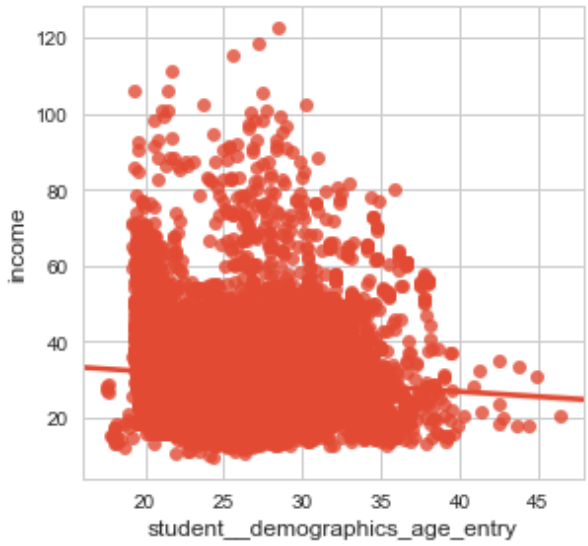
school__online_only	median	mean	count	std	min	max
Distance-education only	37.95	38.292857	28	12.013725	17.2	61.8
Not distance-education only	27.7	29.58269	8440	11.099483	9.4	151.5

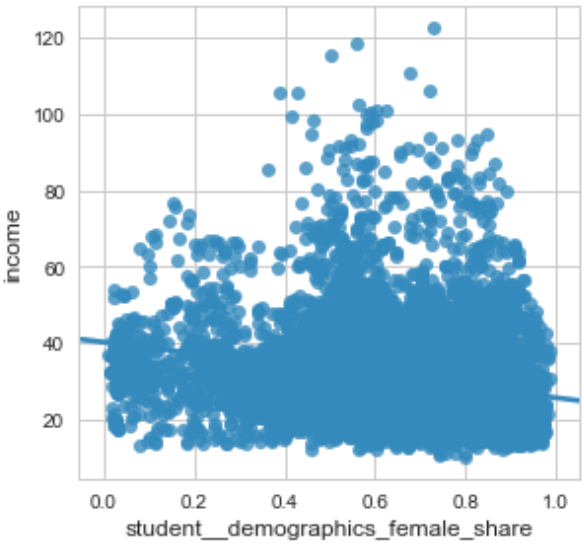
Lastly, the ownership structure has a clear impact on future student earnings where private nonprofit institutions result in the highest average income.

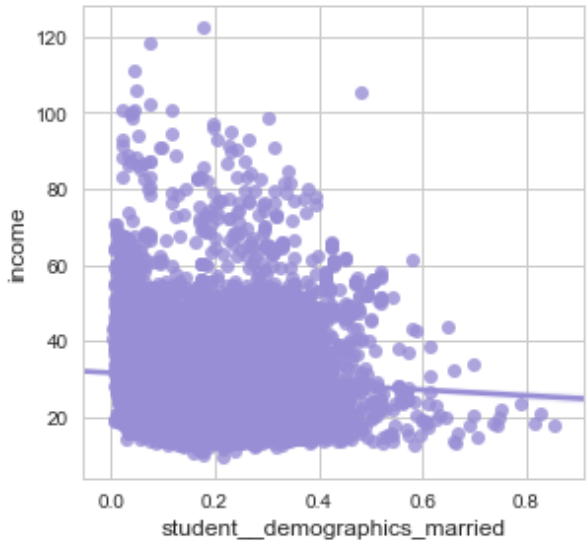
Category 'Student'

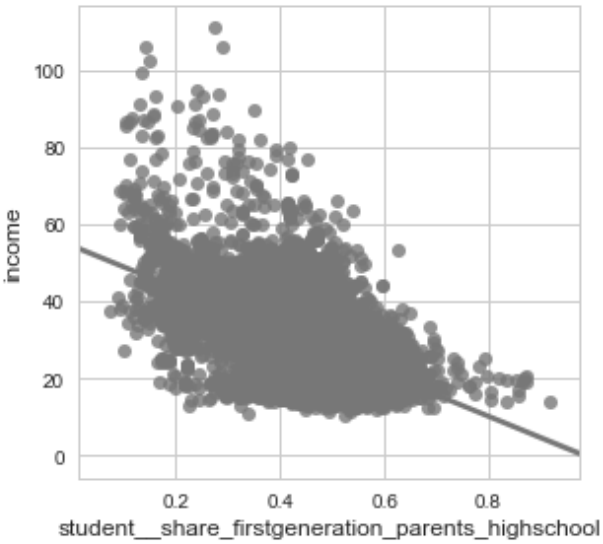
Most obviously, the data reveals that the share of first generation students is negatively affecting income. In the plots below it is apparent that the share of students which parents only completed high school is negatively correlated with income. On the contrary, more students with parents that visited a college mean a higher future average income.

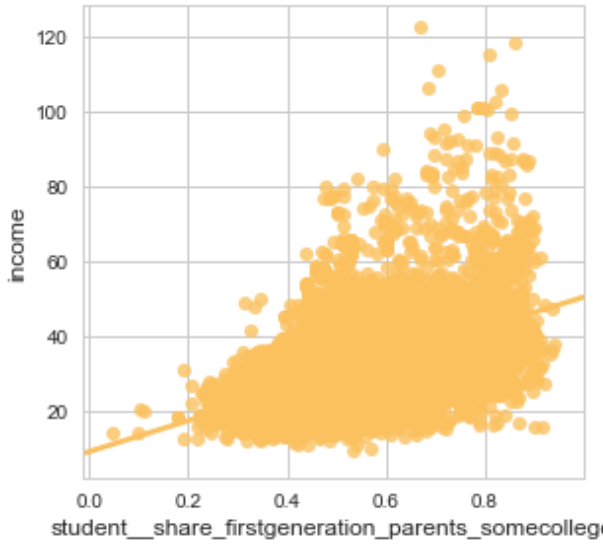
Some other features do not show such strong relationships but are nevertheless interesting such as the age entry, the share of female students and the share of married students. All show a slightly negative correlation with income.

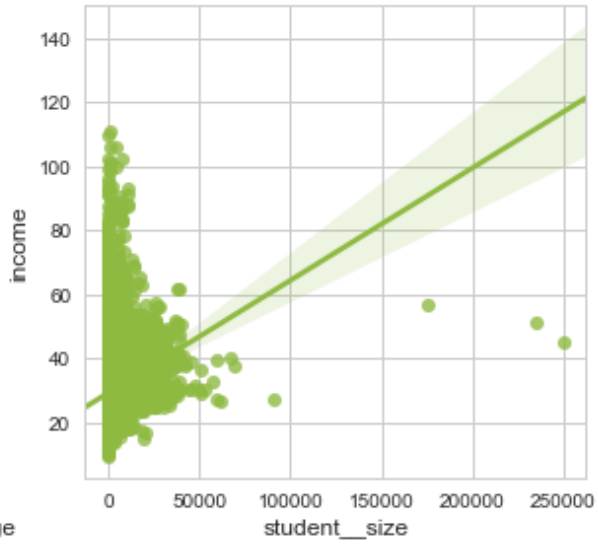

income
120
100
80
60
40
20
20
25
30
35
40
45
student__demographics_age_entry


income
120
100
80
60
40
20
0.0
0.2
0.4
0.6
0.8
1.0
student__demographics_female_share

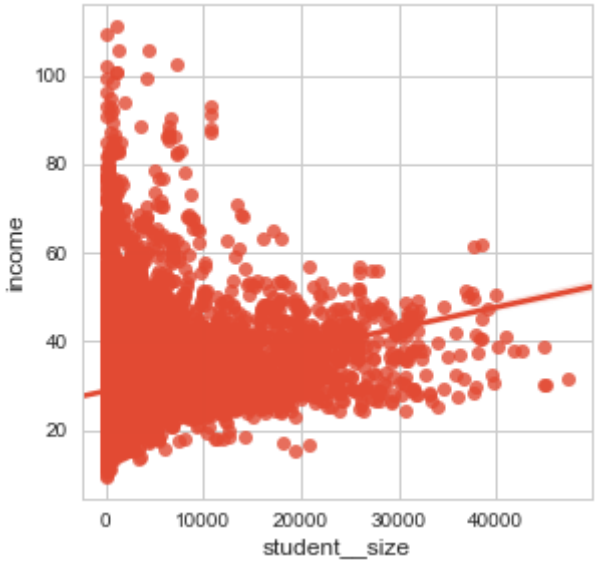

income
120
100
80
60
40
20
0.0
0.2
0.4
0.6
0.8
student__demographics_married


income
100
80
60
40
20
0
0.2
0.4
0.6
0.8
student__share_firstgeneration_parents_highschool


income
120
100
80
60
40
20
0.0
0.2
0.4
0.6
0.8
student__share_firstgeneration_parents_somecollege


income
140
120
100
80
60
40
20
0
50000
100000
150000
200000
250000
student__size

Lastly, some outliers in the plot above heavily distort student size. If we correct for that we see a positive relationship with income:



Regression

For this problem, a Gradient Boosting Regression was chosen, which is a quite robust regressor and does not require too much pre-processing of the data.

Decision trees such as the gradient boosting regression can deal very well with outliers that have been imputed with for instance a very low value. Consequently, all the missing values were imputed with the value -10,000.

Apart from this straightforward manipulation a decision tree does not require any further cleaning of the data as the scale of the features has no impact on the outcome nor have outliers. The same is true for collinearity, as the decision tree will arbitrarily choose one of the correlated features. Lastly, decision trees can also deal with non-linear relationships.

None of the above-mentioned points are true for linear regression. A linear regression would require a lot of careful data cleaning. Hence, the gradient boosting regression was the preferable option.

One last step before running the regression was to create dummies for the categorical values in order to convert them into numerical features.

Consequently, after splitting the data into a train and a test dataframe (70% / 30%), the model was run with the parameters that performed best during various trials. Due to the high number of features, even a randomized search cross-validation required immense computing power, which is why testing of different parameters was done manually.

The best result was achieved with the following parameters:

```
GradientBoostingRegressor(    alpha=0.99,
                              criterion='friedman_mse',
                              init=None,
                              learning_rate=0.05,
                              loss='huber',
                              max_depth=11,
                              max_features='sqrt',
                              max_leaf_nodes=None,
                              min_impurity_split=1e-07,
                              min_samples_leaf=15,
                              min_samples_split=10,
                              min_weight_fraction_leaf=0.0,
                              n_estimators=5000,
                              presort='auto',
                              random_state=5,
                              subsample=1.0,
                              verbose=0,
                              warm_start=False
                          )
```

To score the model the build-in scikit learn score method has been used which returns the coefficient of determination R^2 of the prediction.

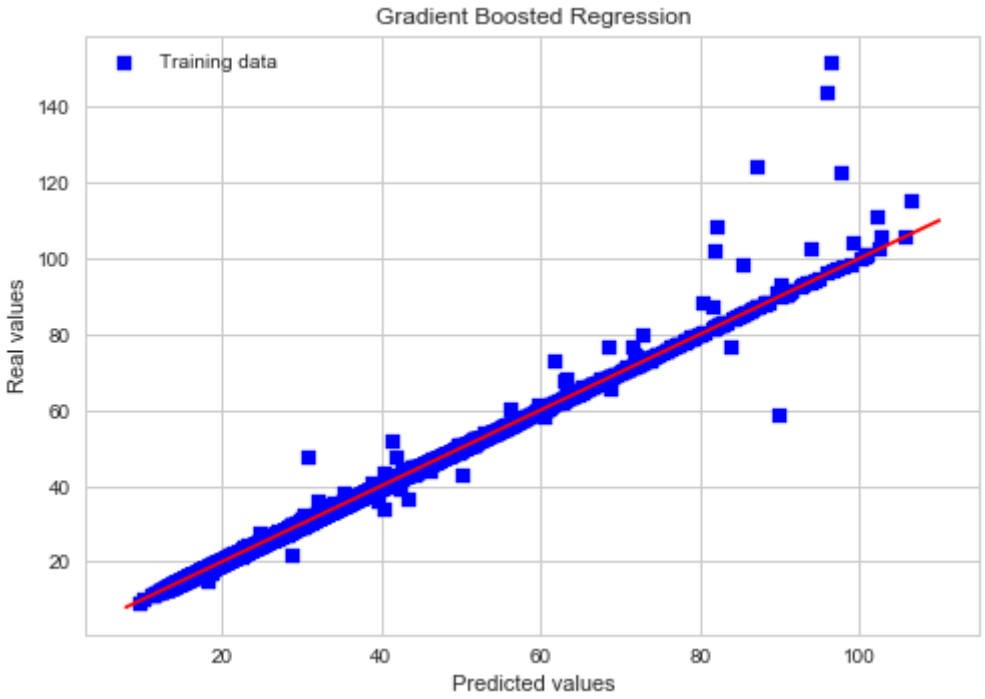
With a best possible score of 1 the model achieved:

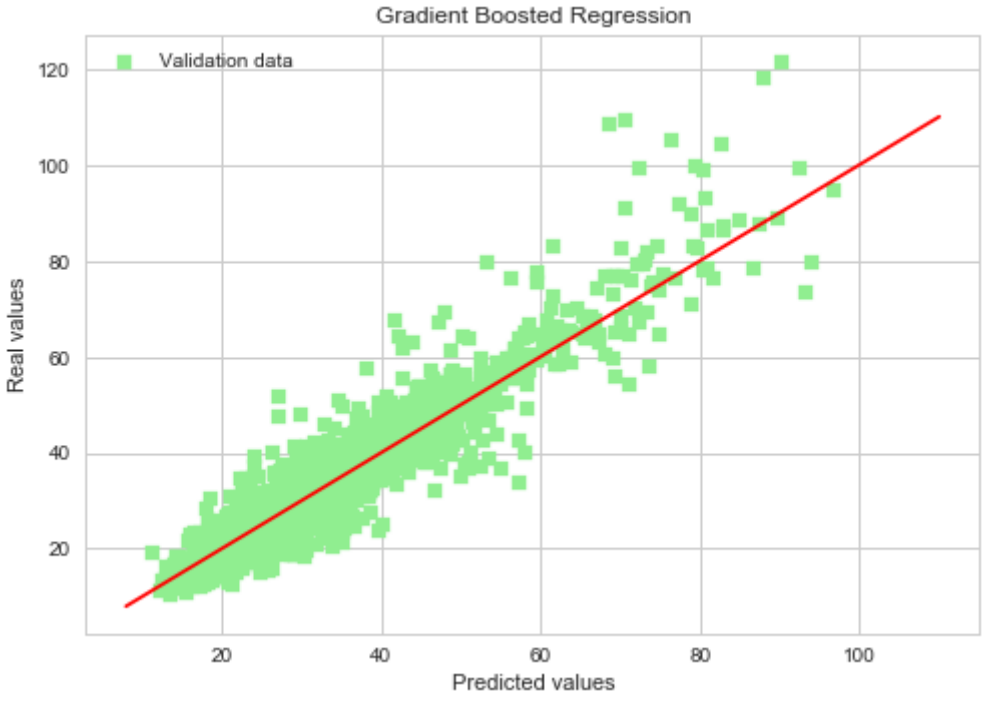
0.901605287421

As a second measure, the root-mean-square error has been calculated:

3.53211744243

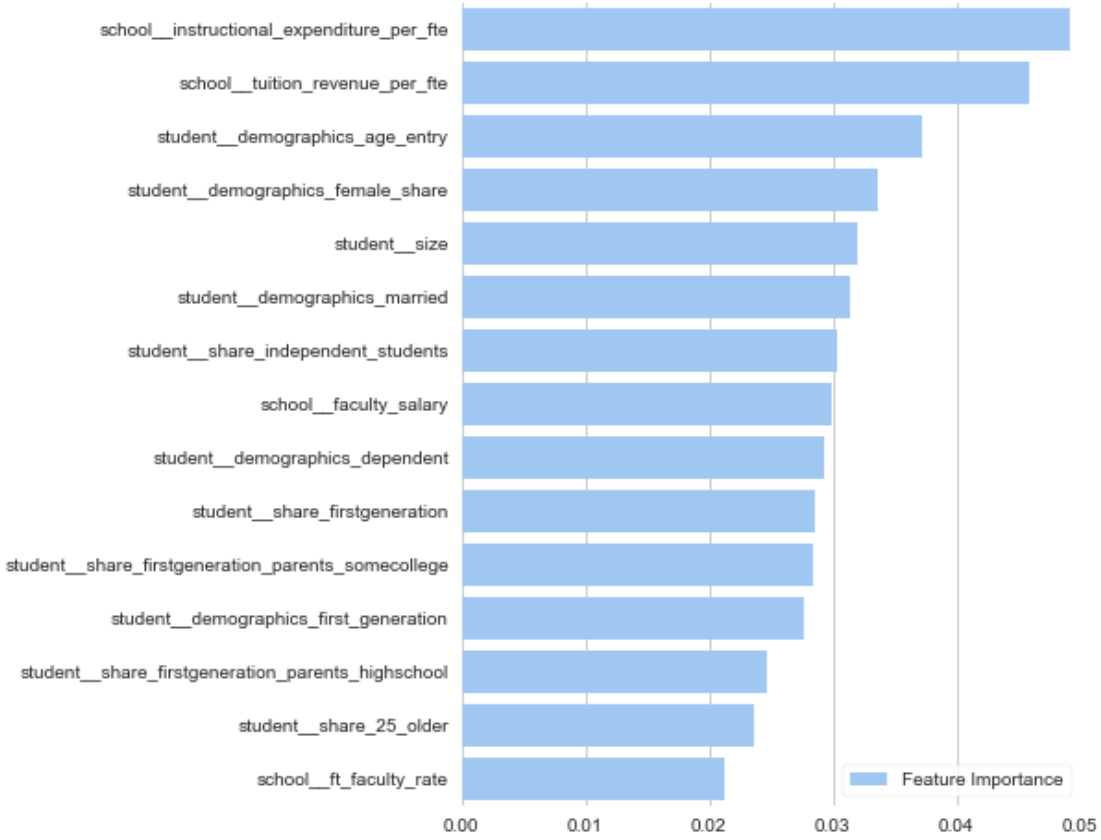
Consequently, it can be stated that the model performs very well which is also proven by the linear relationships between predicted and real values for the training and the testing data:





However, it is also apparent that the model performed much better on the training data. It might be interesting in further studies to look into a potential overfitting of the model. Nevertheless, the model predicts quite accurately.

Also of interest is the importance of the various features. The following chart shows the top 15 features the model deemed most important:



Feature importance is an attribute of the gradient boosting regressor in scikit learn.

After the exploration of the dataset described earlier it does not come as a surprise that the top 15 features only consist of features from the categories ‘school’ and ‘student’.

Conclusion

The model has shown that certain features of the categories ‘school’ and ‘student’ influence future income the most.

For the category ‘school’ it can be concluded that instructional expenditures per full-time equivalent student, net tuition revenue per full-time equivalent student and the average faculty salary are particularly important and positively correlated with future income. Furthermore, the proportion of faculty that is full-time plays a role.

For the category ‘student’ it has been shown that features such as age, size and marriage status are important. Furthermore, all the indicators of first generation students showed up in the analysis of most important features. It appears that the education of the parents plays an important role for students and their potential future income. This might also be related to the variable ‘independent students’ which itself is also listed as an important feature. It is straightforward to assume that students with economically well-doing parents (due to their own academic education) are more independent students.