# College Retention and Completion

**Machine Learning Engineer Nanodegree**

Reynard Hilman
March 19, 2016

# Definition

## Project Overview

College Scorecard (https://catalog.data.gov/dataset/college-scorecard) provides a wealth of data about colleges and university in the US on a yearly basis. The data for 2013 consist of 7804 colleges and universities with more than 1700 features (columns). The features includes data such as SAT and ACT scores, average faculty salary, tuition cost, college type and size, etc.  This project's goal is to find out what factors (features) predict strong retention and graduation rate, and build a model that takes the input of college features and predict its retention and graduation rate. (Source: Udacity's education capstone projects ideas)

## Problem Statement

College retention rate is a number between 0 and 1 which indicates the percentage of students who return to the institution after the first year. Graduation (completion) rate is a number between 0 and 1 which indicates percentage of students that completed the degree within 150% of the expected time to completion. Our model will be a regression model that takes the input of college characteristics and predicts the retention and graduation rate.

There are more than 1700 features to choose to build the model. However, only a few features contain complete data for all colleges. There is not a single college that has all the data for all features. There are colleges without retention or graduation rate. Because retention and graduation rate are the target features for our model, we will first remove colleges that do not have retention and graduation rate. The challenge is to find as many features that affect the retention and graduation rate while minimizing the number of missing data we have to work with. For the remaining incomplete features, if the data is continuous value, we can fill the missing data with the mean, or try to build a model that predicts the missing data. Once we have preprocessed the data and fill in the missing values, we can start building and tuning the model.

## Metrics

To measure how good our regression model, we will use one of the regression metrics such as the R squared.

# Analysis

## Data Exploration

From 1700 features, 48 features that might be relevant are selected to begin with (including the retention and graduation rate). Even from this smaller subset of features, there are still a lot of missing values. To get an idea how complete each feature is, here is the number of available data for each feature:

| | | |
|---|---|---|
| ACTCMMID | ACT | 1342 |
| ADM_RATE_ALL | Admission rate | 2484 |
| AVGFACSAL | Avg faculty salary | 4654 |
| C150_4_POOLED | Completion 4yr pooled | 2472 |
| C150_L4_POOLED | Completion <4yr pooled | 4018 |
| CCBASIC | Carnegie classification-basic | 4355 |
| CCSIZSET | Carnegie classification-Size & settings | 3576 |
| CCUGPROF | Carnegie classification-Undergrad profile | 3559 |
| CONTROL | Control (public/private) | 7804 |
| COSTT4_A | Avg cost academic year | 4137 |
| COSTT4_P | Avg cost program year | 2541 |
| DEBT_MDN | Median debt | 7094 |
| DEBT_MDN_SUPP | Median debt suppressed | 7094 |
| DEP_INC_AVG | Avg income dependent stu | 7580 |
| DISTANCEONLY | Distance only | 7383 |
| GRAD_DEBT_MDN | Median debt complete | 6987 |
| GRAD_DEBT_MDN_SUPP | Median debt completer suppressed | 7094 |
| IND_INC_AVG | Avg income independent stu | 7582 |
| INEXPFTE | Expense per FTE student | 7362 |
| LOCALE | Degree of urbanization | 7380 |
| NPT4_PRIV | Avg net price title IV institut private | 4753 |
| NPT4_PUB | Avg net price title IV institut public | 1923 |
| NUM4_PRIV | Num Title IV student, private | 4785 |
| NUM4_PUB | Num Title IV student, public | 1924 |
| PAR_ED_PCT_1STGEN | % 1st gen students | 7597 |
| PAR_ED_PCT_HS | % parent education high school | 7597 |
| PAR_ED_PCT_MS | % parent education middle school | 7597 |
| PAR_ED_PCT_PS | % parent education post secondary | 7597 |
| PCTFLOAN | % Fed student loan | 7063 |
| PCTPELL | % Pell Grant receiver | 7063 |
| PFTFAC | Full time faculty rate | 4127 |
| PFTFTUG1_EF | Undergrad 1st-time degree seeking | 3686 |

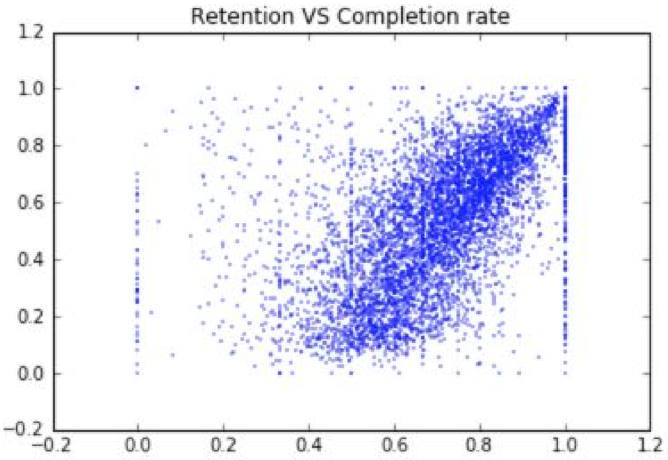
|  |  |  |
|---|---|---|
| PREDDEG | Predominant degree awarded | 7804 |
| RET_FT4 | Retention 4yr | 2348 |
| RET_FTL4 | Retention <4yr | 3920 |
| SATMTMID | SAT math | 1315 |
| SATVRMID | SAT reading | 1301 |
| SATWRMID | SAT writing | 793 |
| SAT_AVG | SAT | 1420 |
| SAT_AVG_ALL | SAT all | 1531 |
| TUITFTE | Net revenue per FTE student | 7362 |
| TUITIONFEE_IN | In state tuition | 4415 |
| TUITIONFEE_OUT | Out of state tuition | 4196 |
| TUITIONFEE_PROG | Tuition fee program year | 2712 |
| UG25abv | % undergrad > 25 yr | 7031 |
| UGDS | Number of Undergrad degree seeking | 7090 |
| WDRAW_DEBT_MDN | Median debt non-completer | 6995 |
| region | Region | 7804 |

There is not a single college that have all the feature above. So we'll have to drop some features and/or fill the missing values with the mean or build an intermediate model that predicts the missing values.

The selected features have 4 different data types:

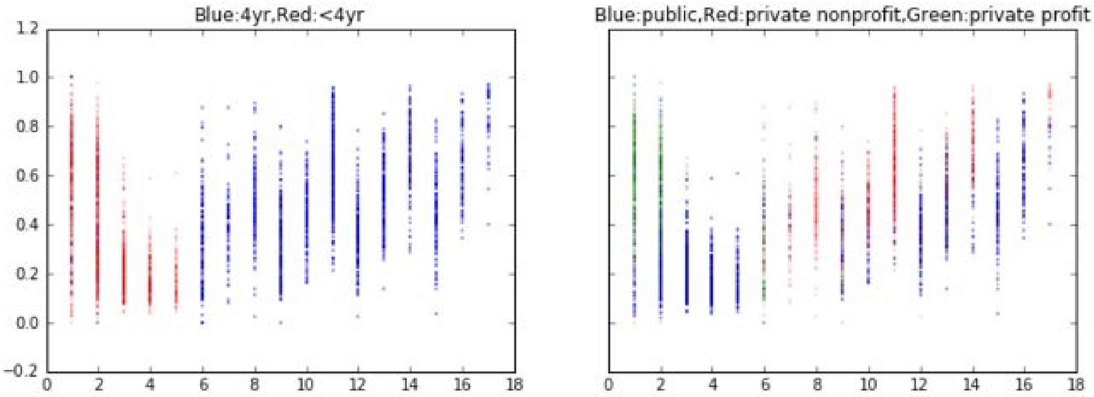
- Percentage data which contains value between 0 and 1
- Monetary data (such as cost, debt)
- Size / count of something (such as number of students)
- Categorical data (such as public, private, region)

## Exploratory Visualization

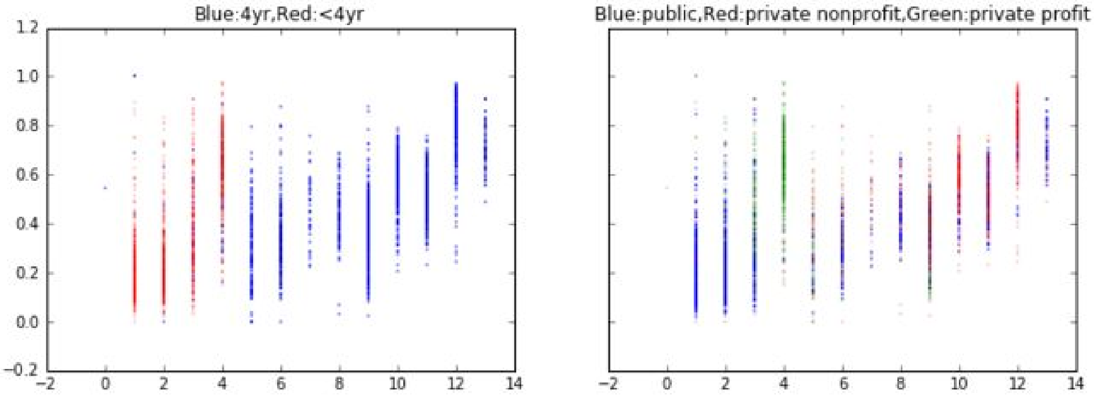


As expected, retention and completion rate have a positive linear correlation. Next, we want to get some idea how the other selected features correlate with the target feature. Because retention and completion rate are linearly correlated, the plot for retention is similar to the completion. To shorten this report, I only include the completion plot here, but the retention plot can be viewed on the exploratory.ipynb file. The plots also show 4 year and <4 year college, as well as public, private and private for profit in different colors.

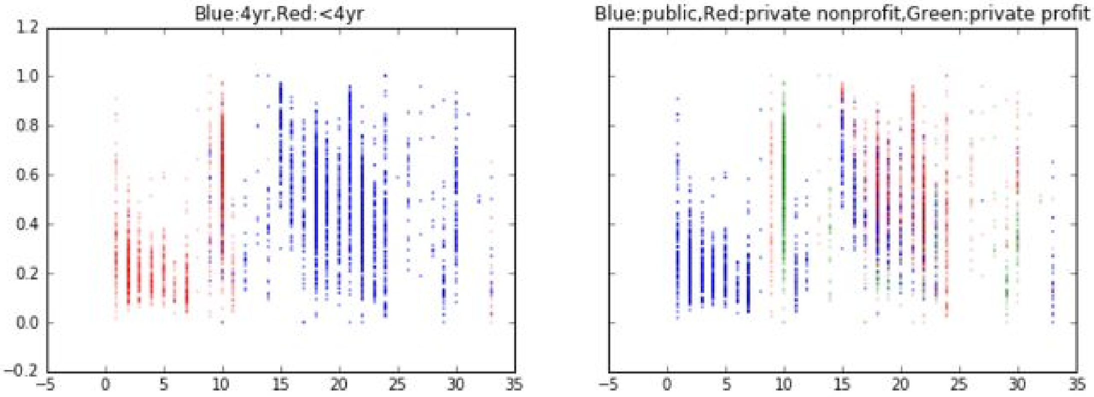
Carnegie classification-Size & settings VS completion



Carnegie classification-Undergrad profile  VS completion

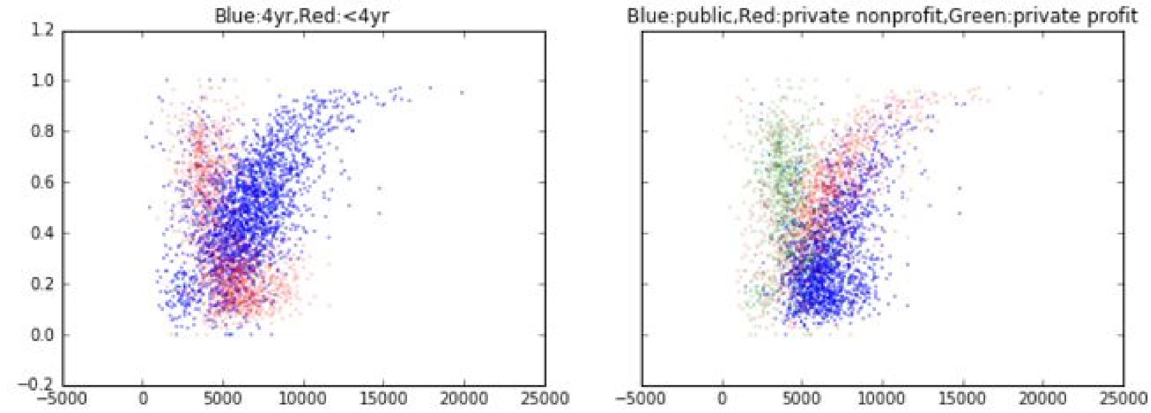


Carnegie classification-basic VS completion

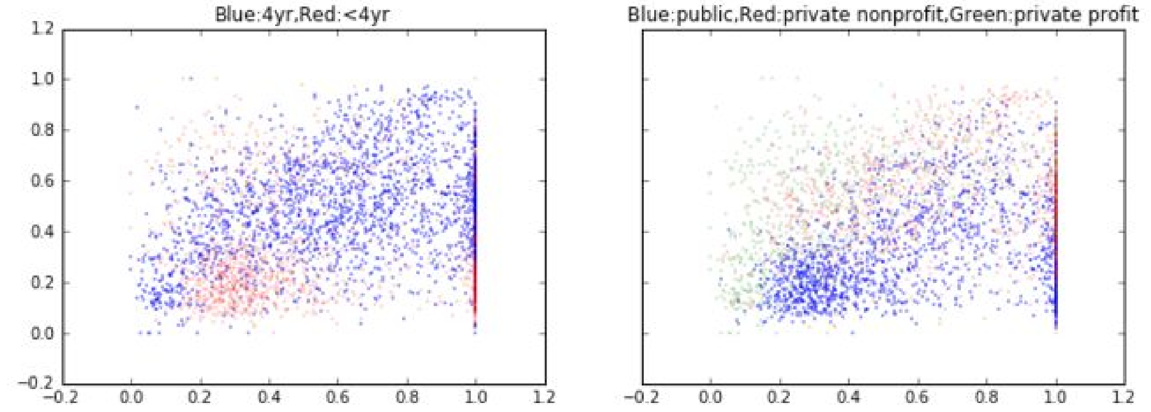


This gives us a pretty good indication that some Carnegie classifications have a higher completion rate than others.

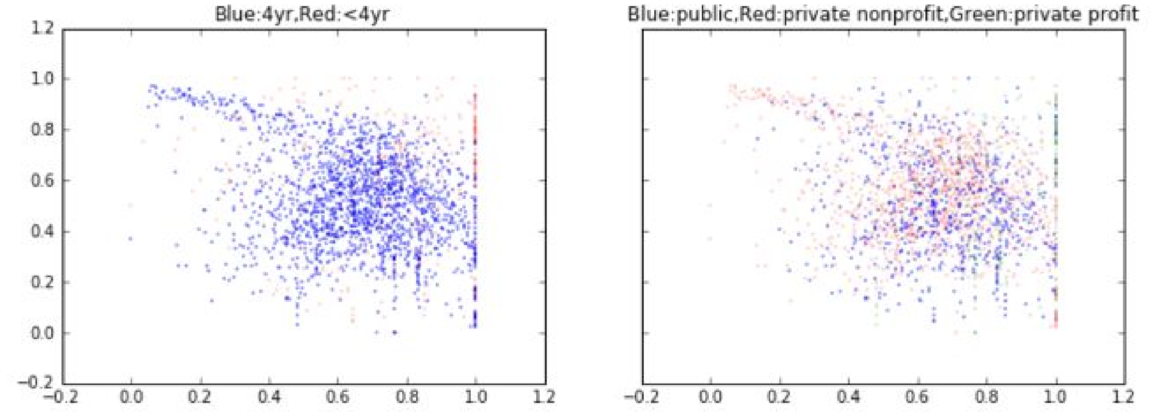
```
Avg faculty salary VS completion
```



```
Full time faculty rate VS completion
```



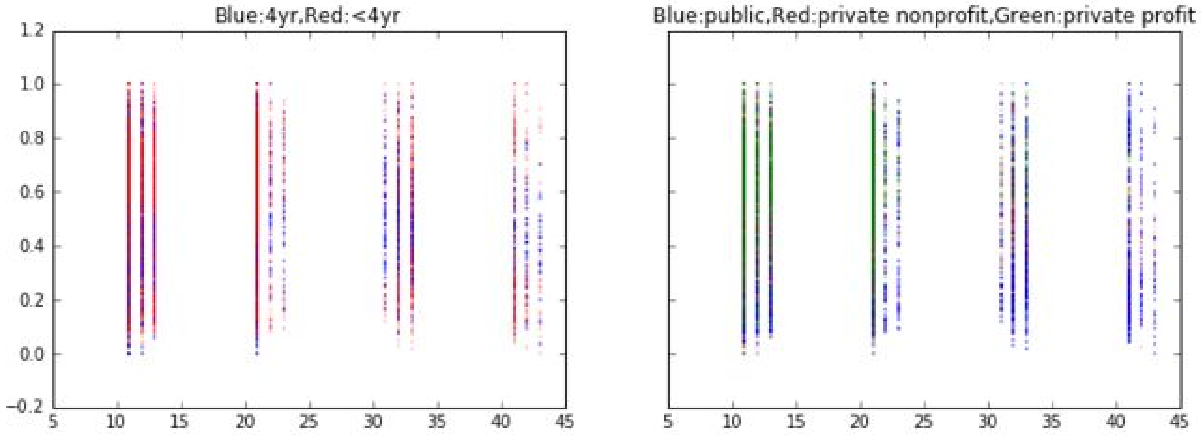
```
Admission rate VS completion
```



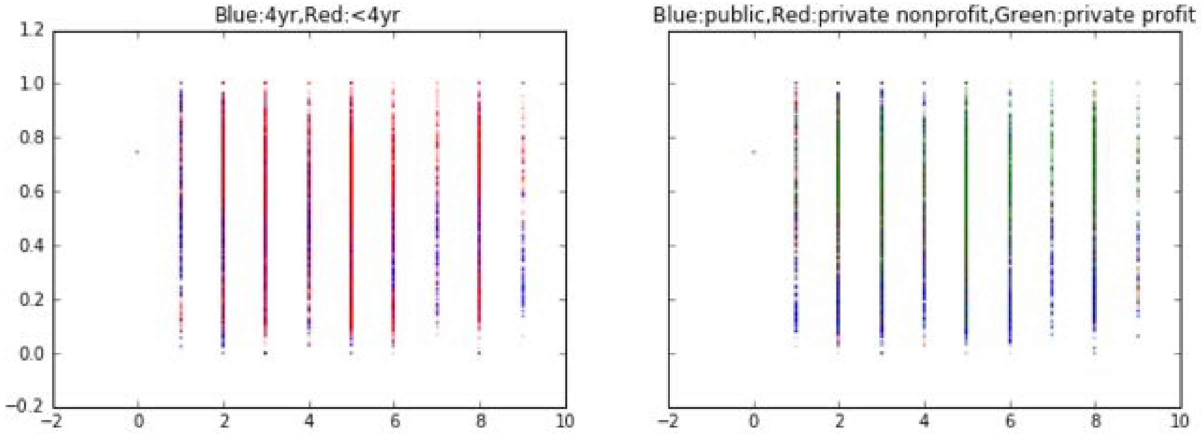
There is a pretty good correlation between average faculty salary and retention especially when we look at it for each different college type. For example for <4 year college, higher faculty salary actually correlate with lower completion rate, but for 4 year college it’s the opposite.

There are some correlations for Full time faculty rate and Admission rate although they are more sparse (varies widely).

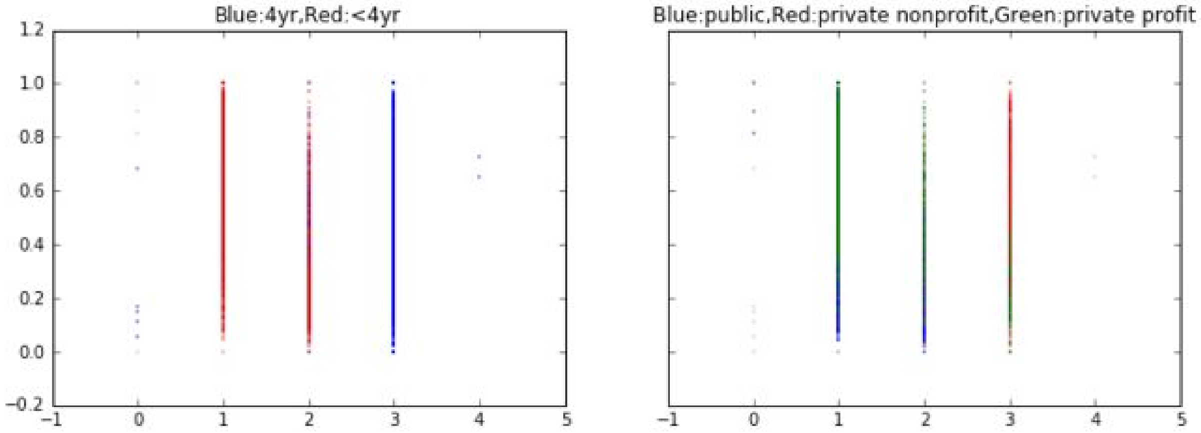
```
Degree of urbanization VS completion
```



```
Region VS completion
```

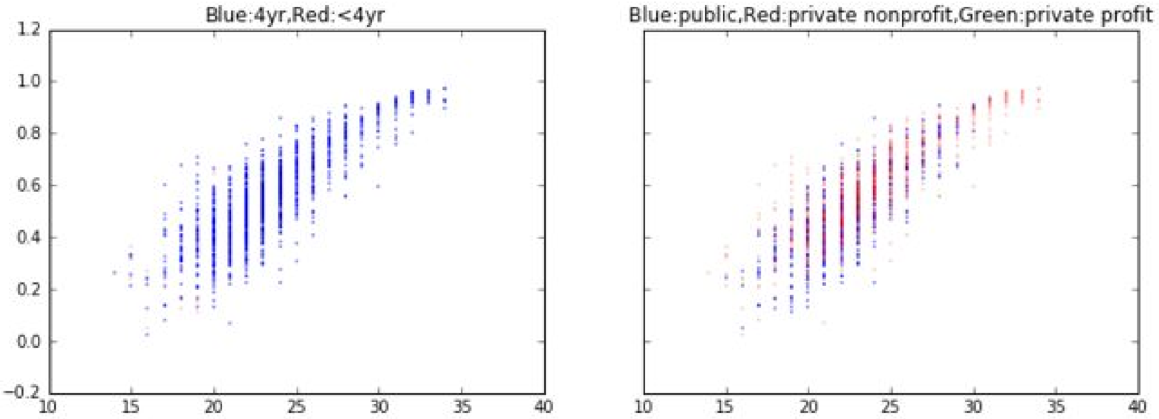


```
Predominant degree awarded VS completion
```

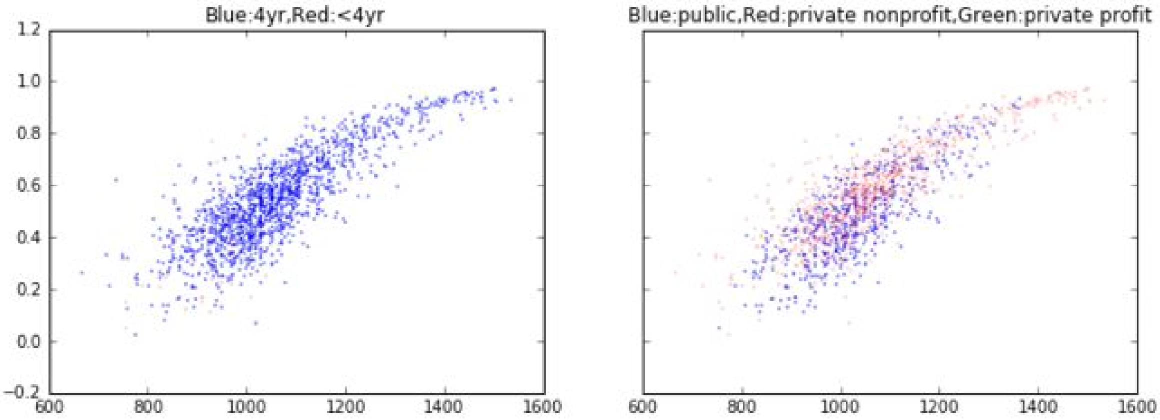


These 3 features (Degree of urbanization, Region, and Predominant degree awarded) do not have much correlation with the completion rate. So these features are a good candidate to be excluded from building the model.

```
ACT VS completion
```

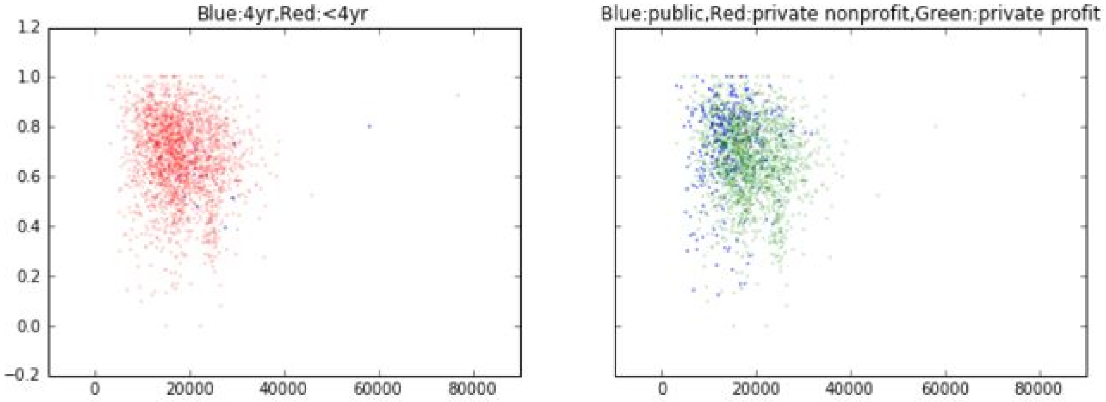


```
SAT VS completion
```

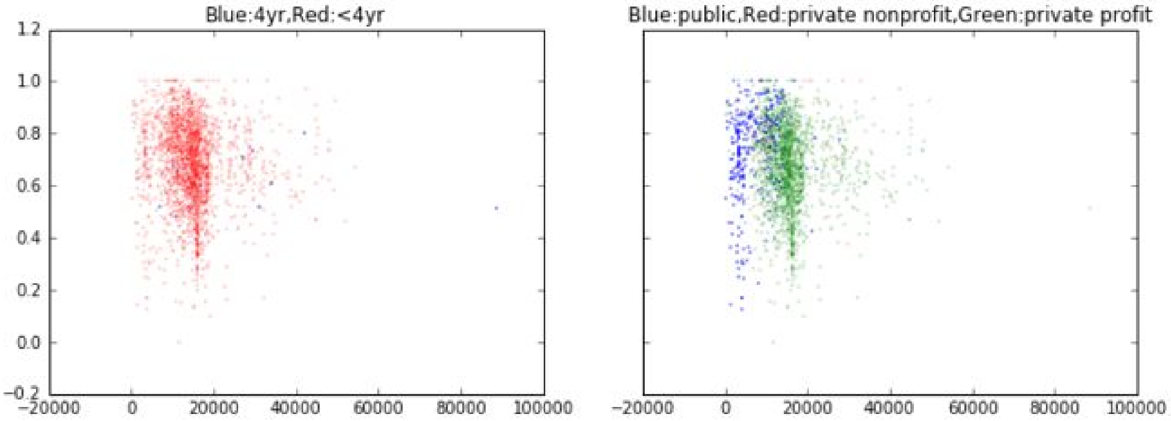


There is a nice correlation between SAT and ACT score with the completion rate.

```
Avg cost program year VS completion
```

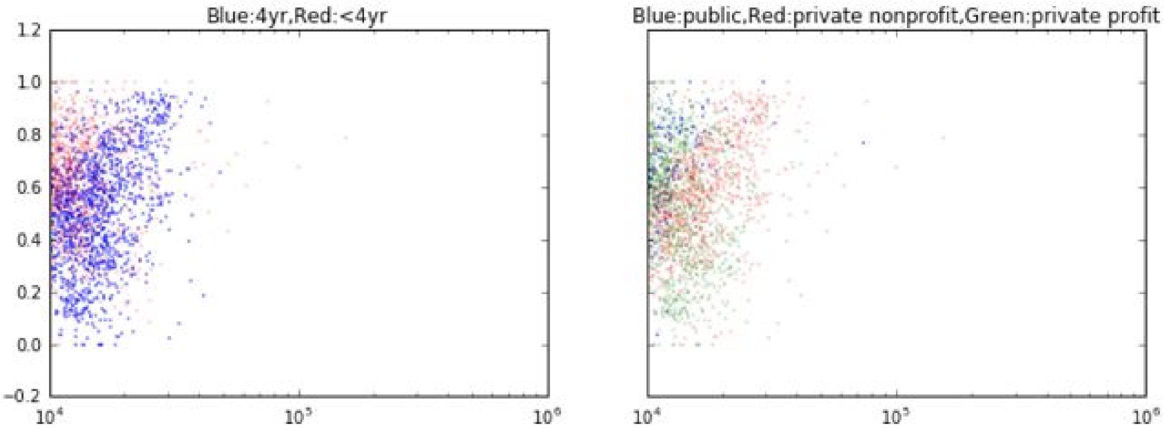


```
Tuition fee program year VS completion
```

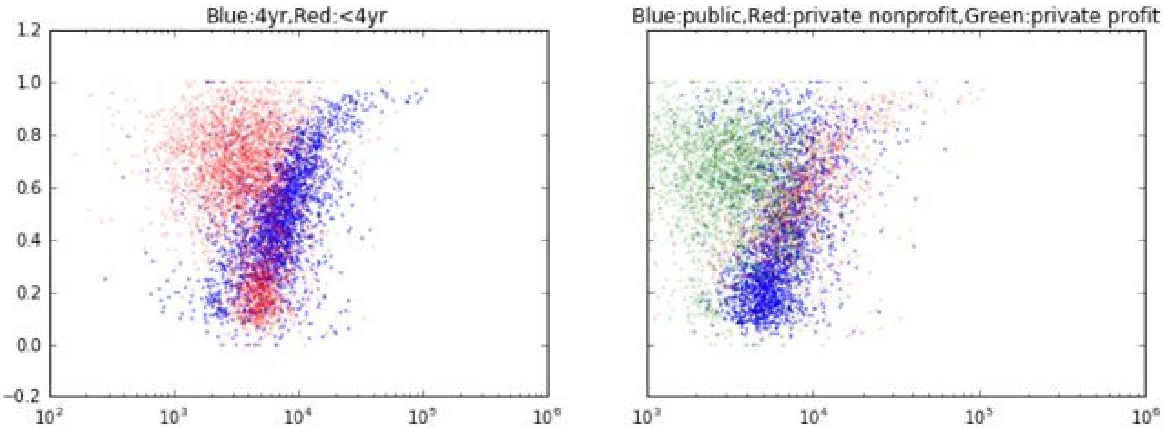


Not very interesting correlation for Average cost program year and Tuition fee program year.

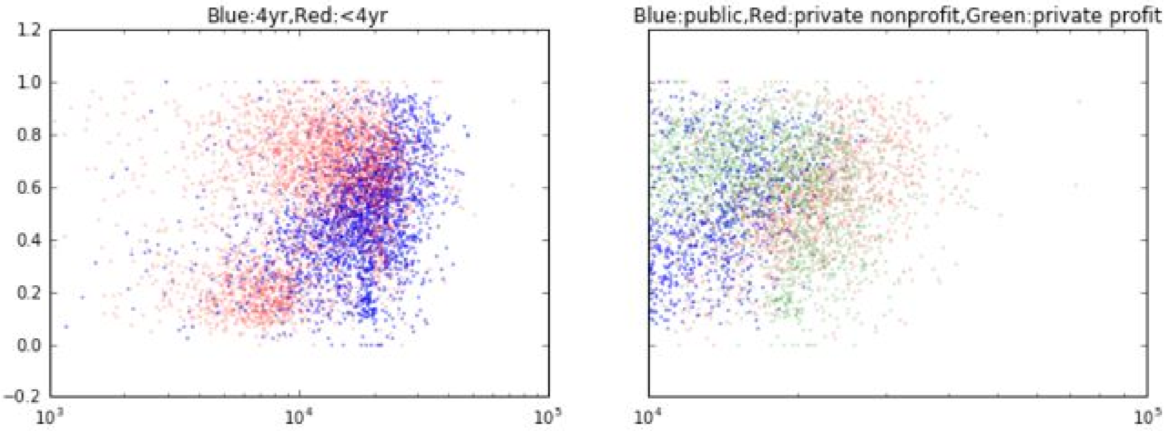
Net revenue per FTE student VS completion



Expense per FTE student VS completion

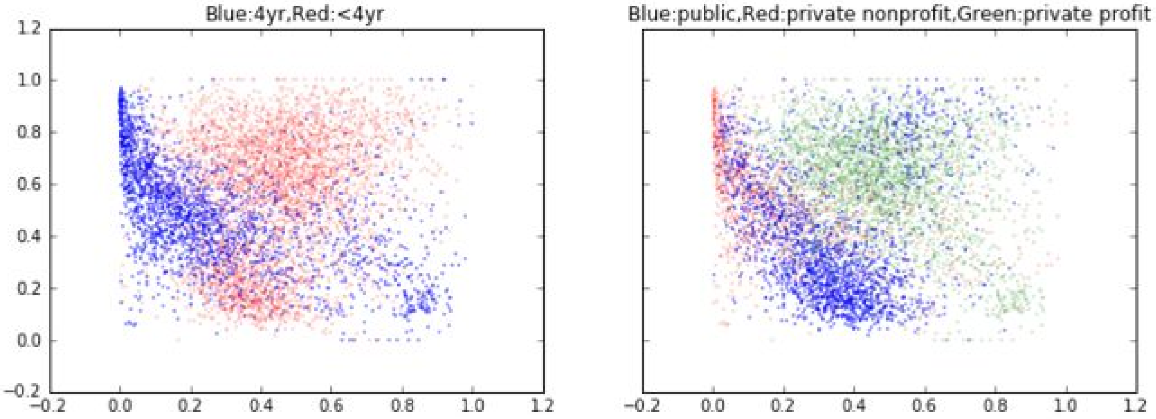


Avg net price Title IV VS completion

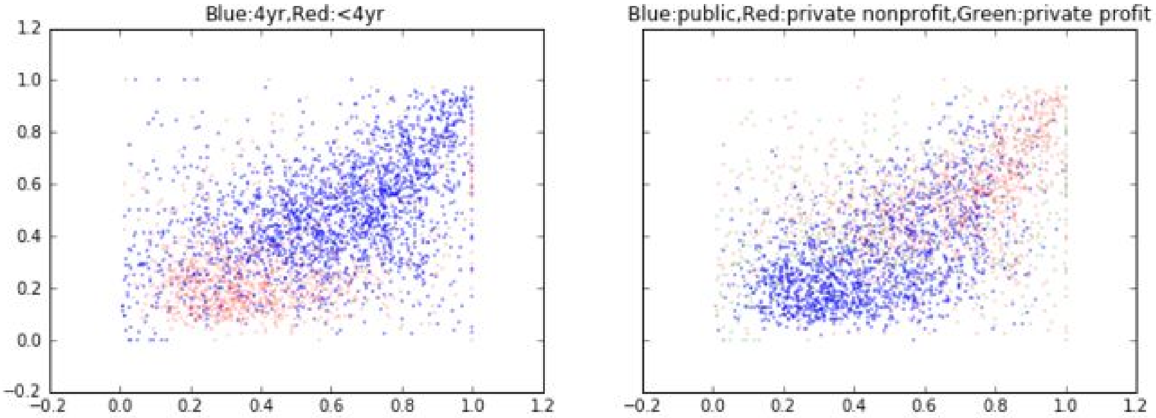


There is some correlation on Expense per FTE student. Not so much for Net revenue per FTE student and Average net price Title IV.

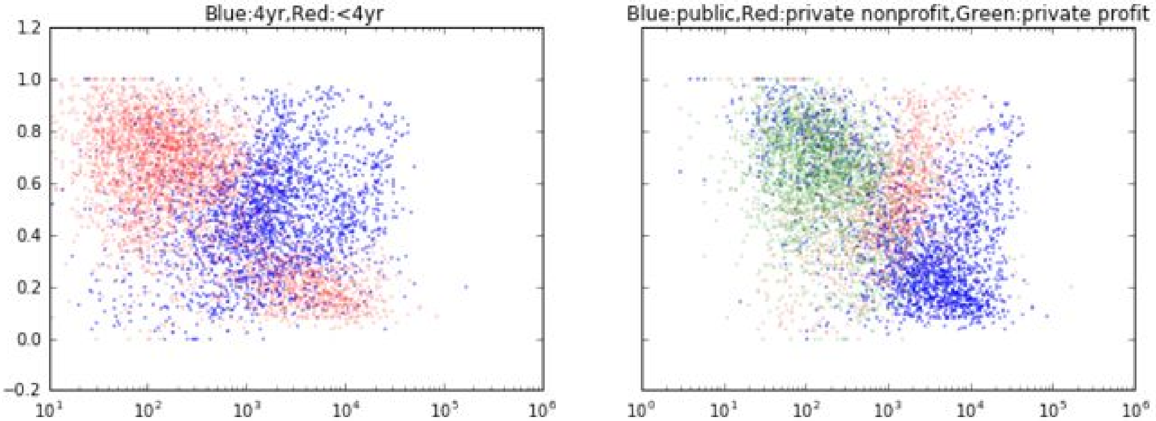
```
% undergrad > 25 yr VS completion
```



```
Undergrad 1st-time degree seeking VS completion
```

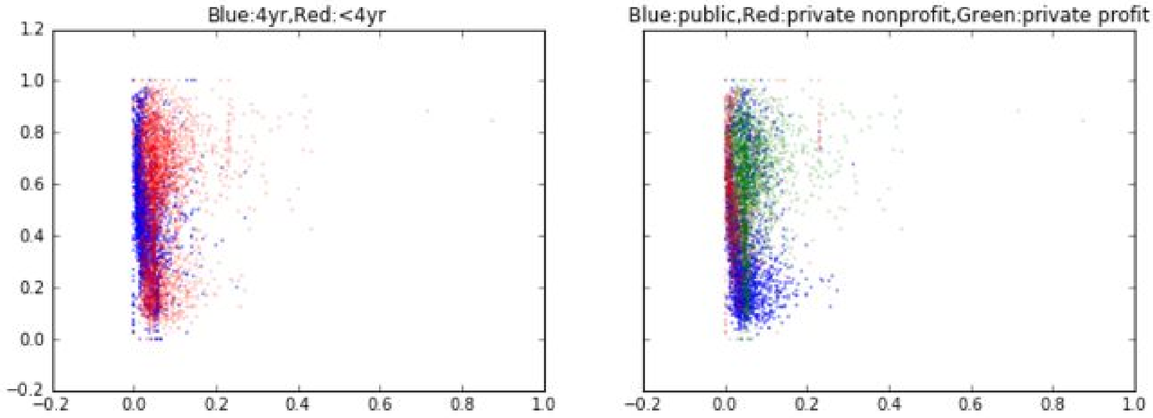


```
Number of Undergrad degree seeking VS completion
```

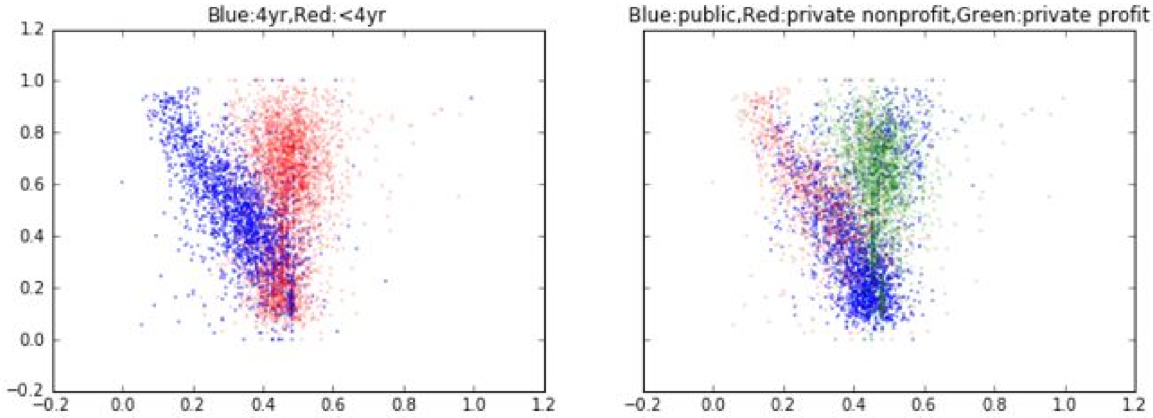


There is definitely a pretty good correlation between Undergrad 1st time degree seeking and completion rate. Four year colleges, public and private colleges have a strong negative correlation between Percentage of undergrad > 25 year and completion rate. There is somewhat a negative correlation between Number of undergrad degree seeking and completion rate, although it’s more sparse.

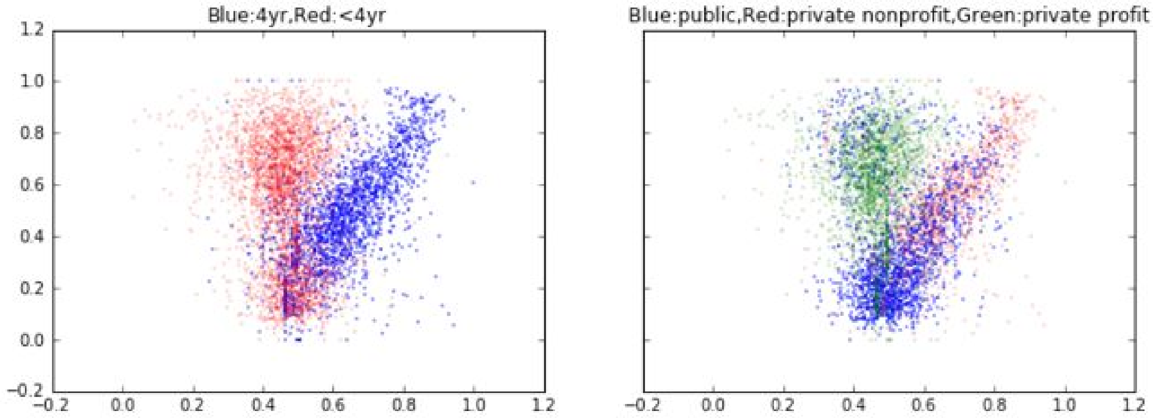
```
% parent education middle school VS completion
```



```
% parent education high school VS completion
```

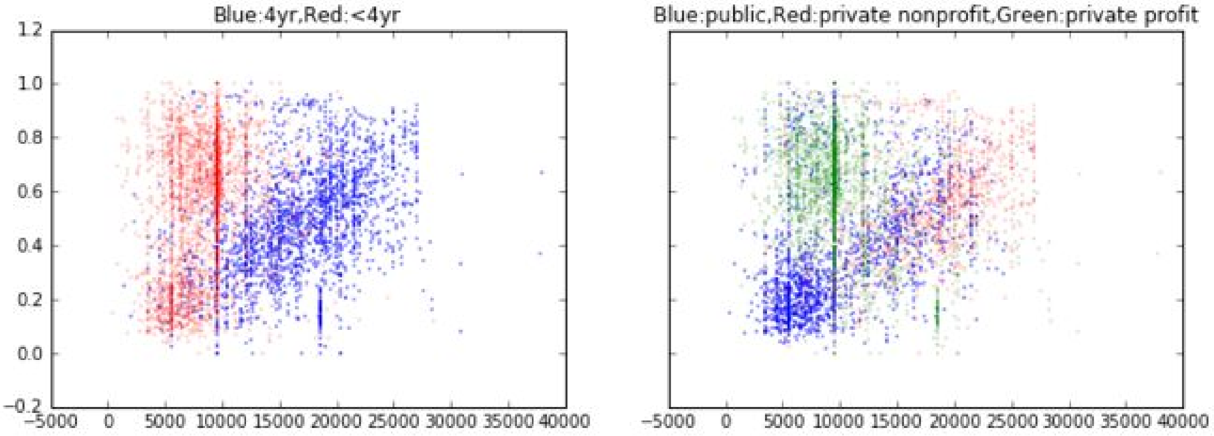


```
% parent education post secondary VS completion
```

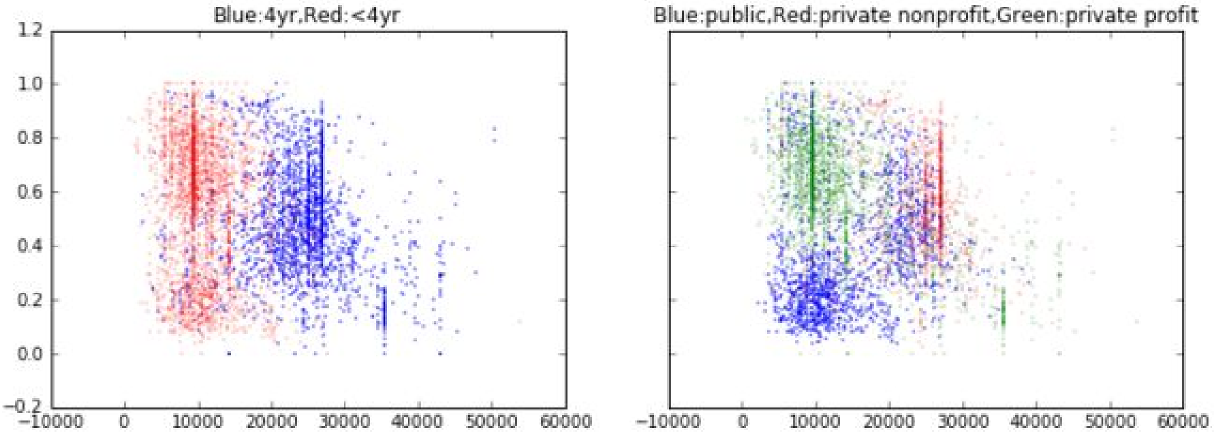


There is a good correlation between Percentage of parent education post secondary and completion rate, and the inverse for Percentage of parent education high school. Percentage of parent education middle school does not have much correlation with completion rate, so that’s another candidate feature to drop.

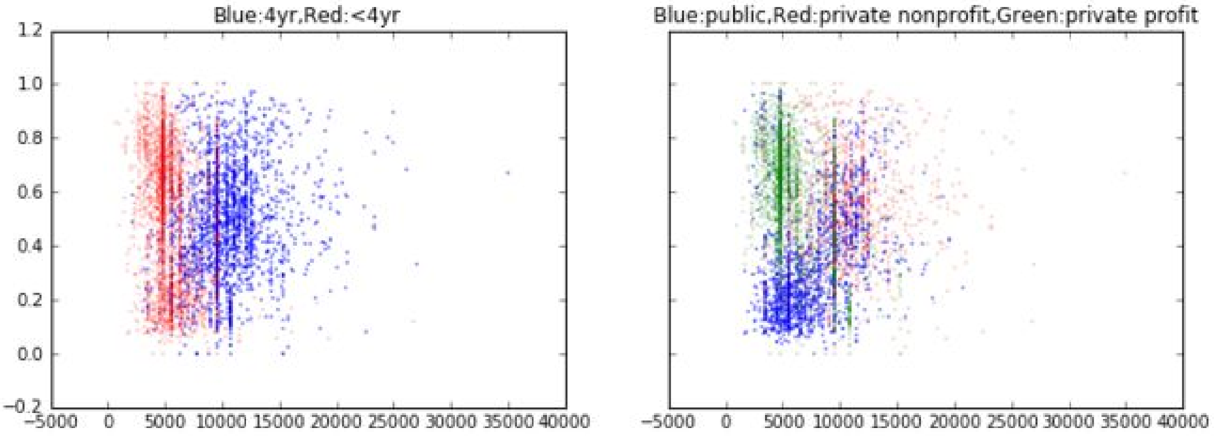
Median debt VS completion



Median debt complete VS completion

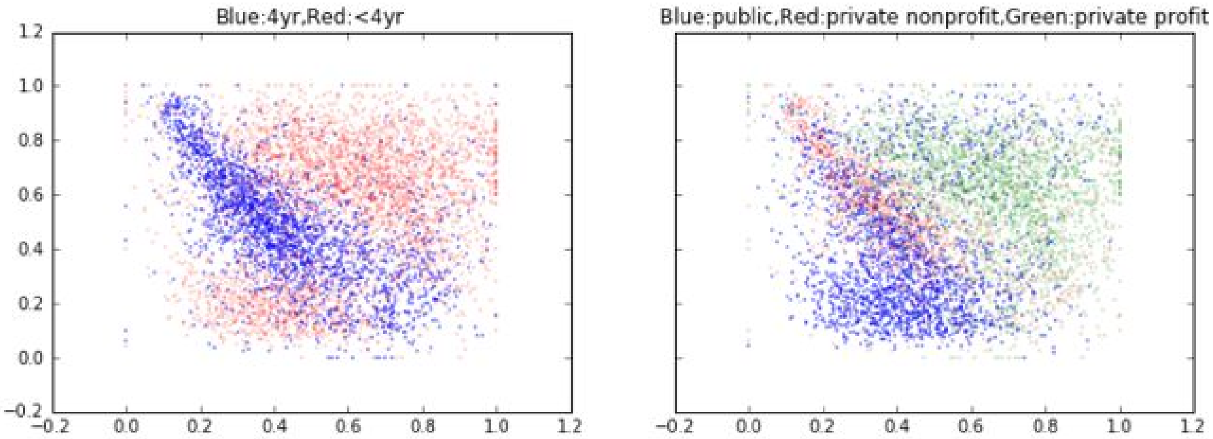


Median debt non-completer VS completion

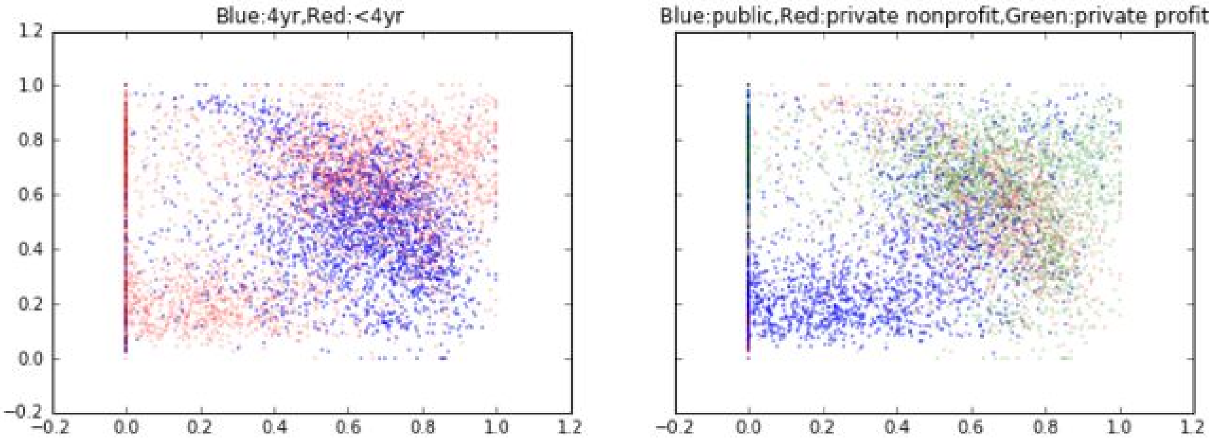


There is definitely some correlation between debt and completion rate. We can also see clusters that separate the different college types.

`% Pell Grant receiver VS completion`



`% Fed student loan VS completion`



Even though at first glance the Percentage of Pell Grant and Fed student loan receiver seem all over the place, there is actually some correlation for some type of college.

## Algorithms and Techniques

Based on the data exploratory analysis, there are some algorithms that might be a good candidate for predicting the retention and completion rate.

- Decision Tree Regressor
  There are some plots that clearly have clusters for the different type of college (such as public, private, 4 year and <4 year). Decision tree should do well in picking these information when splitting the samples.
- SVM Regressor
  Because a lot of the correlation between the features and the target value are pretty linear. Basic SVM (without custom kernel) should do well.

- K-Nearest Neighbor
  This algorithm should do pretty well in predicting the retention and completion rate based on how other similar colleges do.
- Ensemble methods
  If none of the above algorithms perform well enough, we can alway try ensemble method to get a better performance.

Grid search techniques with cross validation will be used to fine tune the parameters.

### Benchmark

We will use the R2 score metric to measure the performance. As a benchmark, a simple decision tree model will be used.

# Methodology

## Data Preprocessing

### Data Cleanup

The original data has 2 completion rate columns which are mutually exclusive. The C150_4_POOLED is for 4 year institution and C150_L4_POOLED is for less than 4 year institution. Because the data from these columns are mutually exclusive, we can combine this into one column and add another boolean column that indicates whether it is a 4 year institution. In the same way retention data (RET_FT4 and RET_FTL4) are also split into 2 columns that can be combined. Some other features that can be combined into one column are NPT4_PRIV, and NPT4_PUB, as well as NUM4_PRIV, and NUM4_PUB for private and public college.

After combining the mutually exclusive columns for retention and completion, we need to get rid of rows that do not have completion or retention rate because those are the target features. We also get rid of rows that have 0 retention rate but high completion rate and vice versa, since that seems more like an anomaly. After this cleanup there are still 5930 rows to work with.

### Missing value treatment

Because there are a lot of missing values, we'll fill the missing values with the mean of the feature. For example, there are only about 1500 colleges with SAT score, so we'll fill the missing SAT scores with the average of the SAT score from the 1500 colleges.

### Data Scaling and Transformation

Based on the 4 different data types in the data exploratory section, we can transform the non categorical features using standard scaler so they have 0 means. The categorical features will be transformed into one hot encoding.

### PCA

PCA analysis on the 23 continuous (non-categorical) features shows that the first 15 PCA components explains more than 96% of the variance in the data. To reduce dimensions and noise, the first 15 PCA components are used to replace the original non-categorical features.

## Implementation

After data preprocessing is done, we split the data for train/test, and save it in a pickle file so that we have a consistent test set. Keeping the test set consistent eliminates one random variable, so we can see if a change in the model affects the result.

4 different algorithms are implemented:

- Decision Tree
- SVR
- KNN
- RandomForest

For each algorithm that we want to explore, we will build 2 regression models. One for predicting the completion rate and one for predicting retention rate. The models will be fed the same preprocessed data as the input. A helper method is implemented for each algorithm. For example:

```
build_SVR_model(X_train, X_test, y_train, y_test, cv=3, params=None)
```

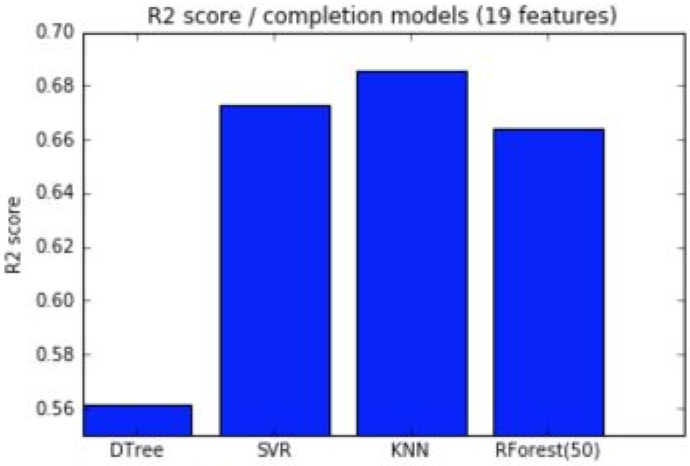
The y_train and y_test are Nx2 arrays containing the completion and retention data. This helper method also accept number of cross validation and params for doing Grid Search. Each of this helper function will return 2 regression models, and report the r2 scores for each model as well as its best params from grid search.

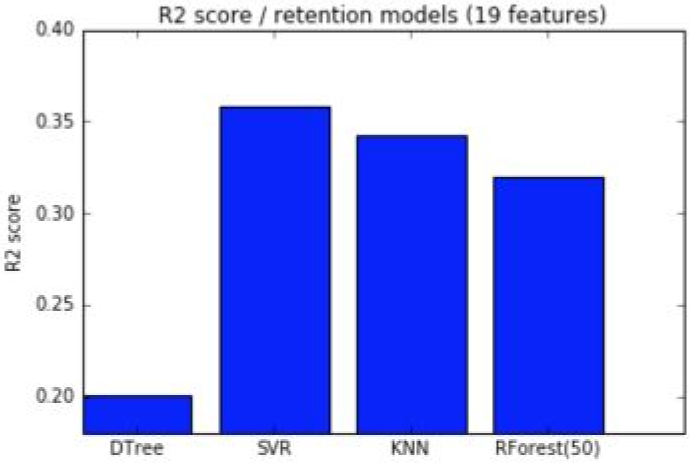
The preprocessed data consist of:

- 15 PCA components.
- 3 one-hot-encoded features for the type of college (public, private non-profit, and private for-profit)
- 1 feature to indicate that it's a less than 4-year college
- 15 one-hot-encoded features for Carnegie classification-Size & settings
- 14 one-hot-encoded features for Carnegie classification-Undergrad profile
- 33 one-hot-encoded features for Carnegie classification-basic

In the first implementation, for each algorithm we build the model using the first 19 columns (15 PCA components, college type and less than 4-year college, without the Carnegie classification features).

The the completion model performs well enough (with R2 ~ 0.68), but the retention model performs much worse in general (with R2 ~ 0.35 at best).



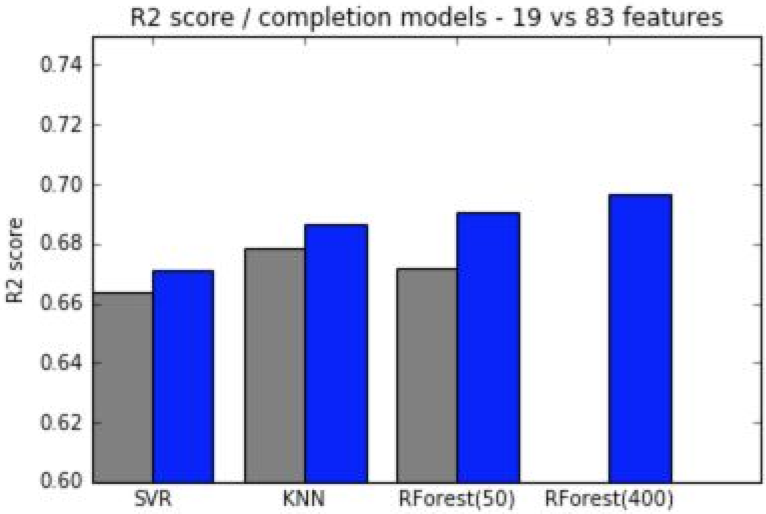


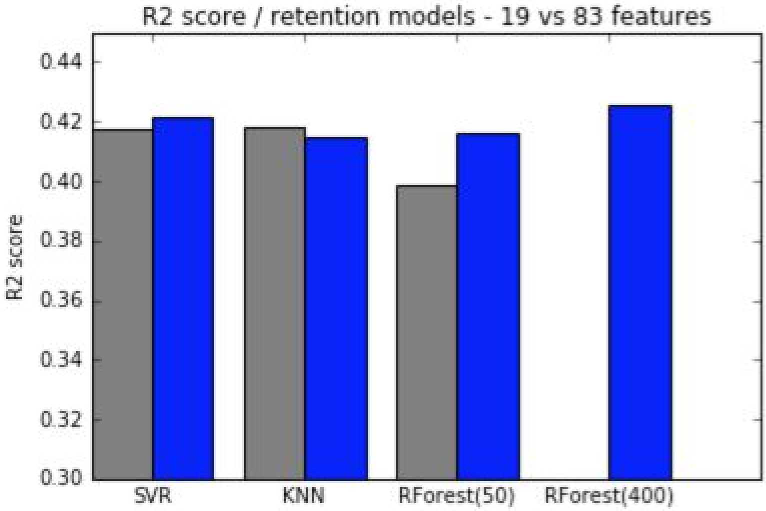
Although it seems that KNN is best for the completion model and SVR for retention, depending on the train/test randomization the result could be different.

## Refinement

The first refinement is to add more features and try RandomForest with more estimators. All 3 Carnegie classifications are added which result in 83 features total. This new input is fed to the same helper methods that build the models. A random forest with 400 estimators is added into the mix as well. The result of adding more features is just a slight performance increase most of the time. Although there are cases (depending on the train/test randomization) where more features perform slightly worse. Random forest with 400 estimators perform slightly better most of the time but there are also cases where it does worse (on the test data). Here is one performance comparison between models with 19 and 83 features:

Grey - models with 19 features
Blue - models with 83 features



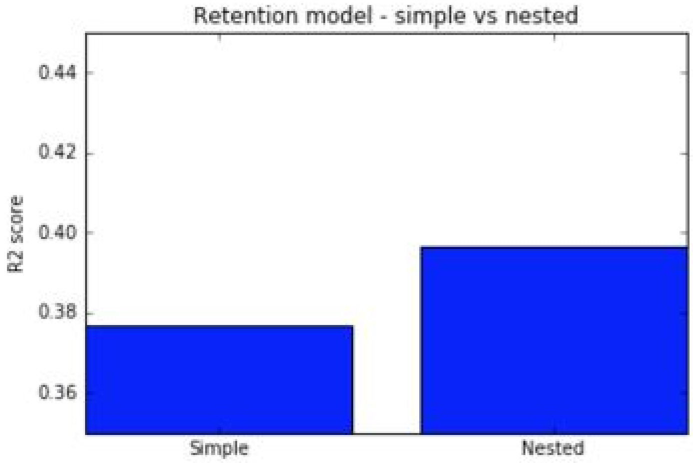


## Using Completion model to help predict Retention

Because the Completion model has a pretty good r2 score compared to the Retention model, we could use the Completion model to help predict the Retention. This might work because there is a clear positive linear correlation between completion and retention.

For this improvement, the implementation for the completion model is exactly the same. However, the retention model will have one more feature--the completion rate. The retention model is trained using the same X_train + y1_train (the completion rate data). When doing prediction on the test set, the retention model takes the X_test + the predicted completion rate from the completion model.

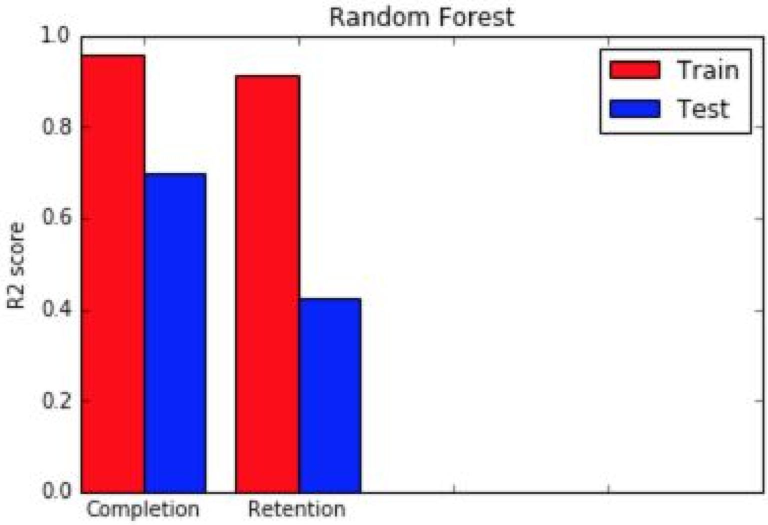
In some cases, the result is a better r2 score for the retention model (as much as 0.02 increase). But depending on the train/test randomization, sometimes the nested model does not improve or even performs worse on the test data.



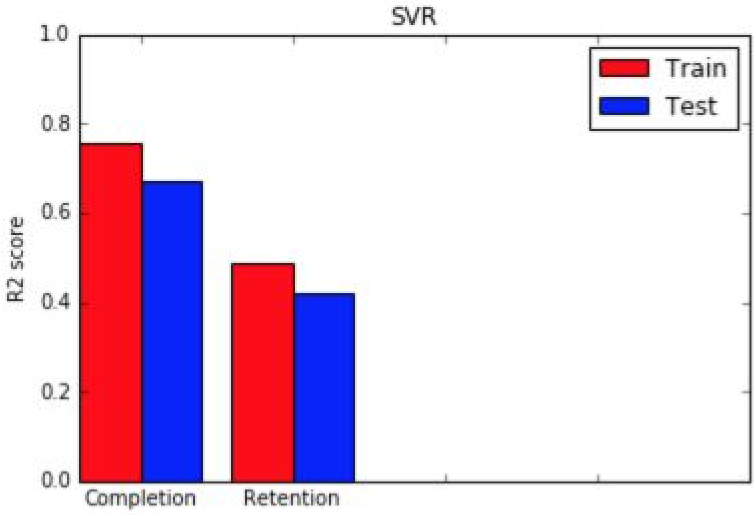
# Results

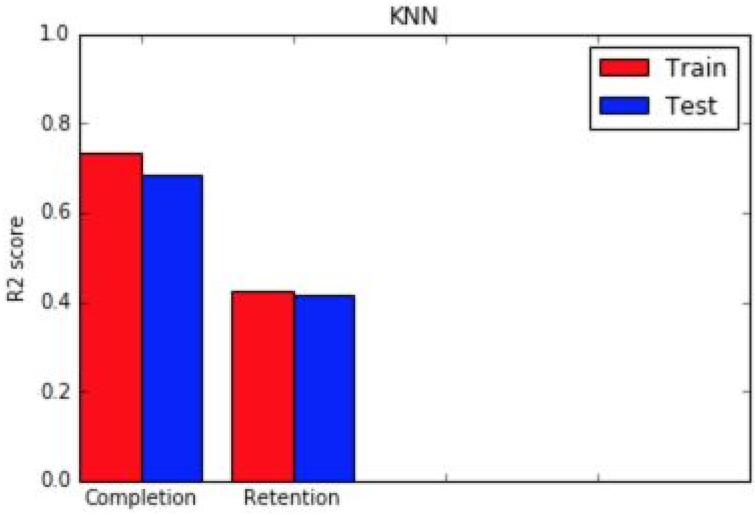
## Model Evaluation and Validation

Even though the three models (SVR, KNN and RandomForest) have similar R2 score for both Completion and Retention prediction, each model has different characteristics. RandomForest model clearly overfits the training data, with a big gap between training and testing R2 score.



SVN and KNN have training and testing R2 score that converges better:

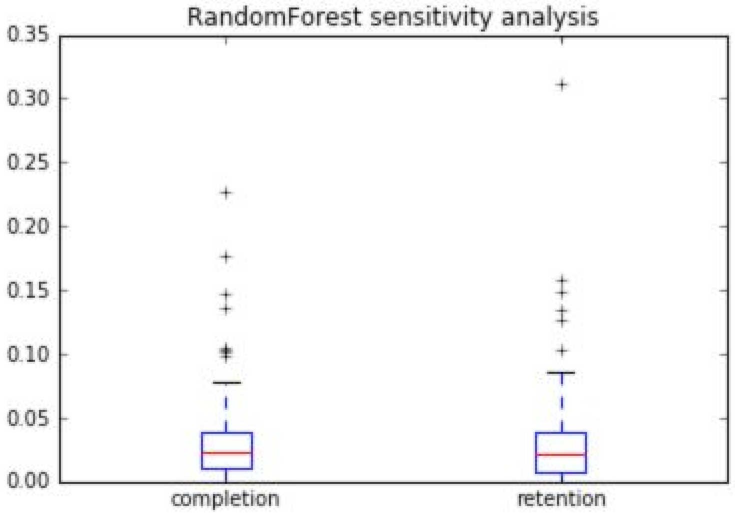


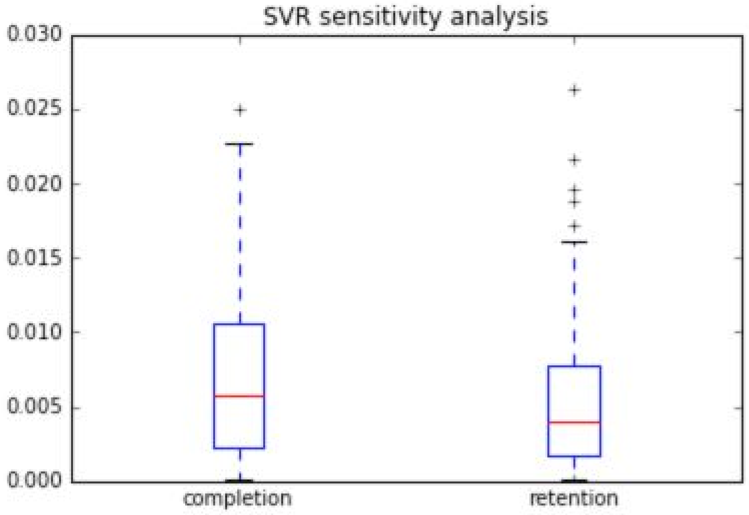


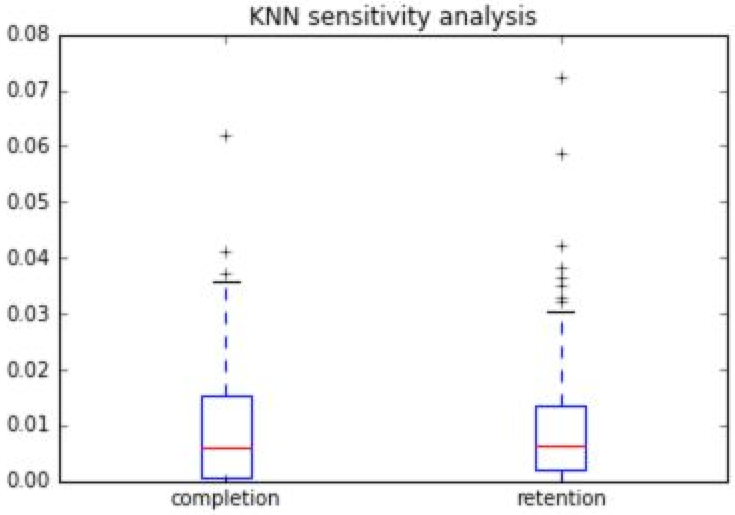
## Sensitivity Analysis

Sensitivity analysis is done by choosing 100 random data from the training set and changing the input by a little bit to see how different the prediction result for each model is. The difference in prediction result are collected and visualized using whisker plot to see the stability of the model.

The result shows that SVR is the most stable. It has the lowest overall difference in prediction result (mean, standard deviation, range). Random Forest is the most erratic model. Some Random Forest prediction differs by as much as 0.35 when the input is changed just by a little bit.


RandomForest sensitivity analysis
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
completion
retention


SVR sensitivity analysis
0.030
0.025
0.020
0.015
0.010
0.005
0.000
completion
retention


KNN sensitivity analysis
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
completion
retention

The final model chosen is the SVR that uses 83 features with the following parameters (for both completion and retention):
epsilon = 0.1
C =  0.1
gamma = 0.1

A few reasons SVR was chosen:

- It consistently performs well enough. It does not always perform the best (when compared with KNN and Random Forest), but it never performs much worse than KNN and Random Forest either.
- The R2 score between train and test data converges pretty well for both completion and retention model.
- SVR is the most stable model on sensitivity analysis.

### Justification

The final SVR model has R2 scores of about 0.67 and 0.42 for completion and retention model respectively. That is quite an improvement from the benchmark (simple decision tree model) R2 scores of 0.61 and 0.27.

# Conclusion

### Reflection

One interesting problem in this project is the fact that there are a lot of missing data. There are so many missing data that there is no single row that has all the data, even from the selected potential features. I tried finding a subset of potential features where we can still have a substantial number of reduced dataset that has all the features.

On model-reduced-dataset.ipynb I was able to reduce the dataset to 1210 rows that have all the data for a subset of the features. The SVR models from this reduced dataset actually get much better R2 scores for both completion and retention prediction (0.82 and 0.7 respectively).

### Improvement

One of the biggest challenge in this project is missing data. The fact that we can build models that perform much better when we reduce the dataset to only those rows that have complete data, means that if we can predict the missing data better, there is a chance that we can build a better model. So rather than filling the missing data with just the mean, we can try building intermediate models that predict the missing data.

Using the model from the reduced dataset, we can conclude that a better model is possible. We can use the R2 scores from the reduced dataset model as a target benchmark, and use the final model from this project as a new baseline benchmark. A better models should have R2 scores between the new baseline benchmark and the target benchmark.