College Retention and Completion

Machine Learning Engineer Nanodegree

Reynard Hilman
May 15, 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

There are a few potential regression error metrics for this problem: R2, Mean Squared Error (MSE) and Median Absolute Error (MAE). MAE will be used to measure how good our model is, because:

- It better communicates what the error means for our model (compared to MSE). For example it’s easier to understand when we say the average prediction error for our model is 0.1, as opposed to saying the average prediction squared error of the model is 0.02.
- In this case, MAE gives better insight on the Retention model’s performance. The R2 score for Retention model happens to be much worse than the Completion model, however the MAE for Retention is actually lower (better) than that of Completion.

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
```

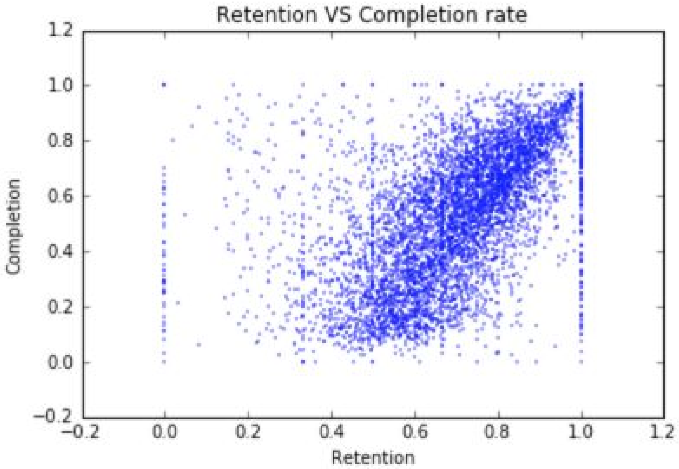
PAR_ED_PCT_1STGEN	% 1st gen students	7597
PAR_ED_PCT_HS	% parent education high school	7597
PAR_ED_PCT_MS	% parent education middle school	7597
PAR_ED_PCT_PS	% parent education post secondary	7597
PCTFLOAN	% Fed student loan	7063
PCTPELL	% Pell Grant receiver	7063
PFTFAC	Full time faculty rate	4127
PFTFTUG1_EF	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:

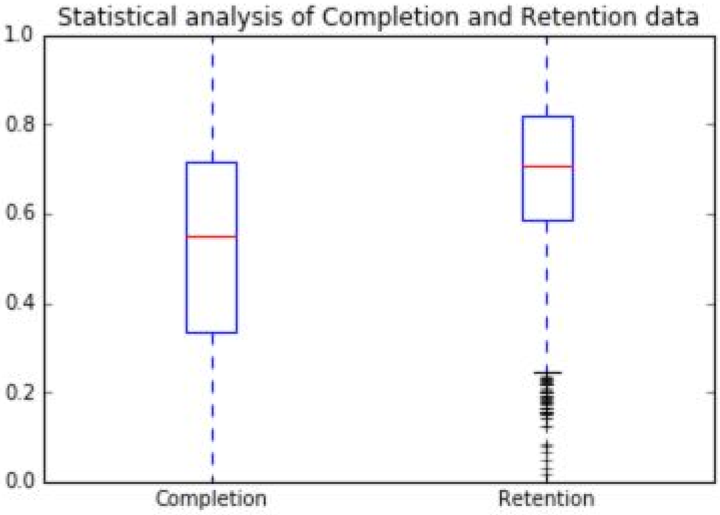
- Percentage data which contains value between 0 and 1
- Monetary data (such as cost, debt)
- Size / count of something (such as number of students)
- Categorical data (such as public, private, region)

Exploratory Visualization



As expected, retention and completion rate have a positive linear correlation. In term of distribution, the completion data is more spread out between 0 and 1 while retention data is more concentrated in 0.5 - 1 range, with a lot of outliers below 0.3.

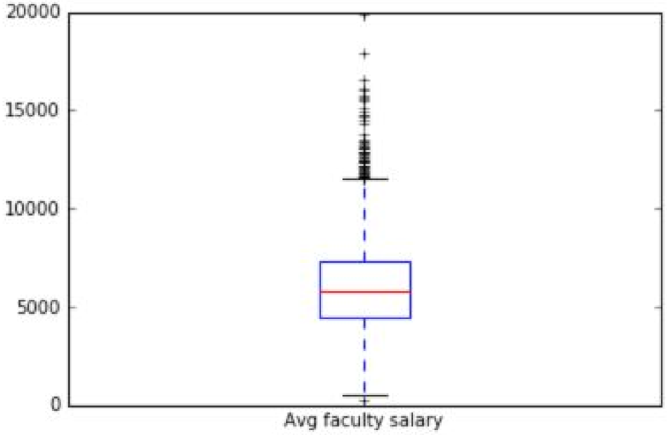
```
--- Completion ---
count    6007.000000
mean        0.526462
std         0.237548
min         0.000000
25%         0.333350
50%         0.551705
75%         0.717280
max         1.000000
--- Retention ---
count    6007.000000
mean        0.692658
std         0.178662
min         0.000000
25%         0.587750
50%         0.707400
75%         0.818200
max         1.000000
```

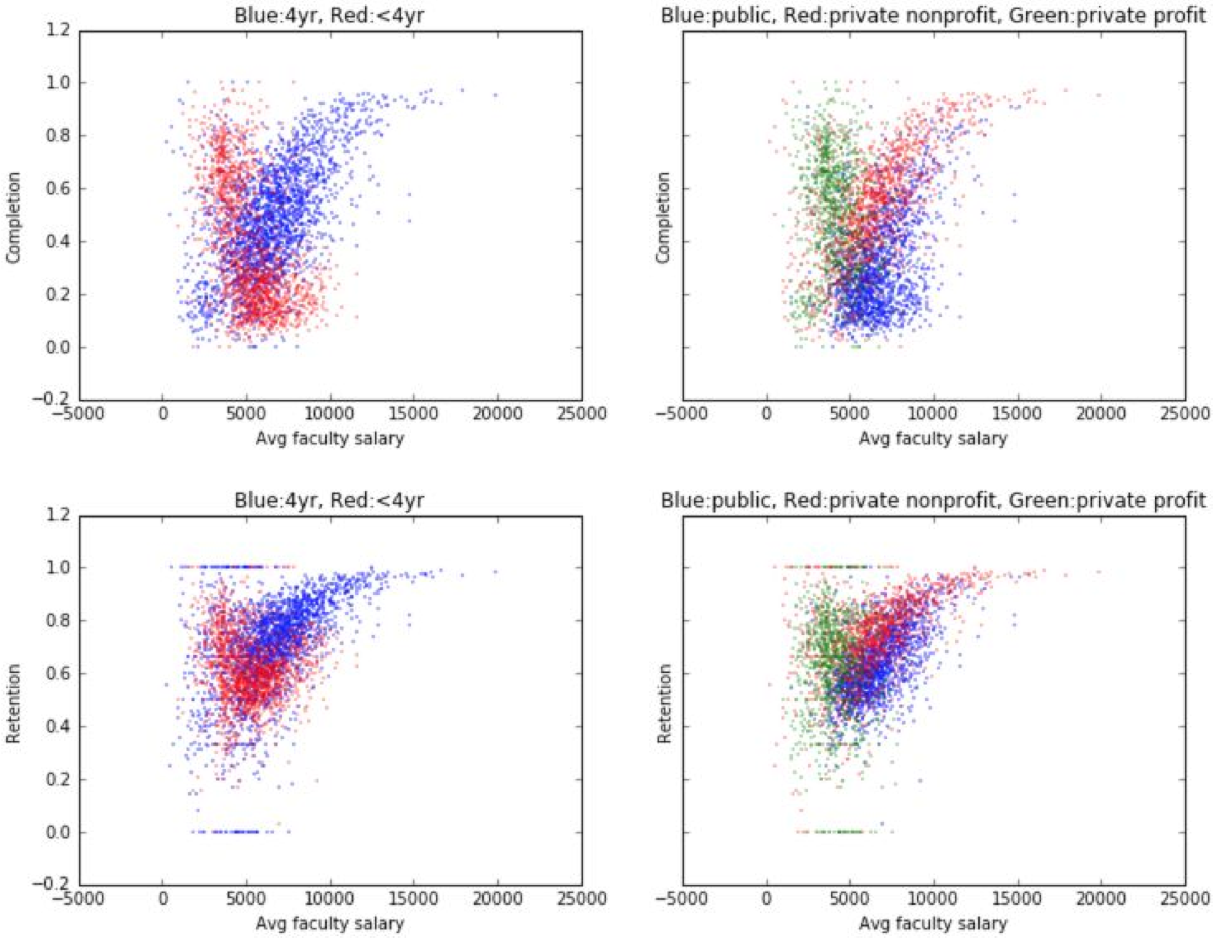


Next, we want to get some idea how the other selected features correlate with the target features. In the exploratory.ipynb, all the potential features are plotted against Completion and Retention. To shorten this report, I only include some features that are interesting here. The plots also show 4 year and <4 year college, as well as public, private and private for profit in different colors.

Average Faculty Salary

```
count     3794.000000
mean      6019.977596
std       2220.556513
min        269.000000
25%       4490.500000
50%       5832.000000
75%       7299.000000
max      19862.000000
Name: AVGFACSAL, dtype: float64
```

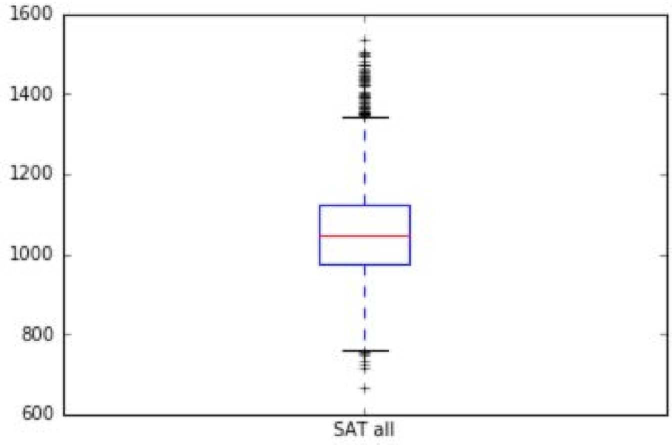


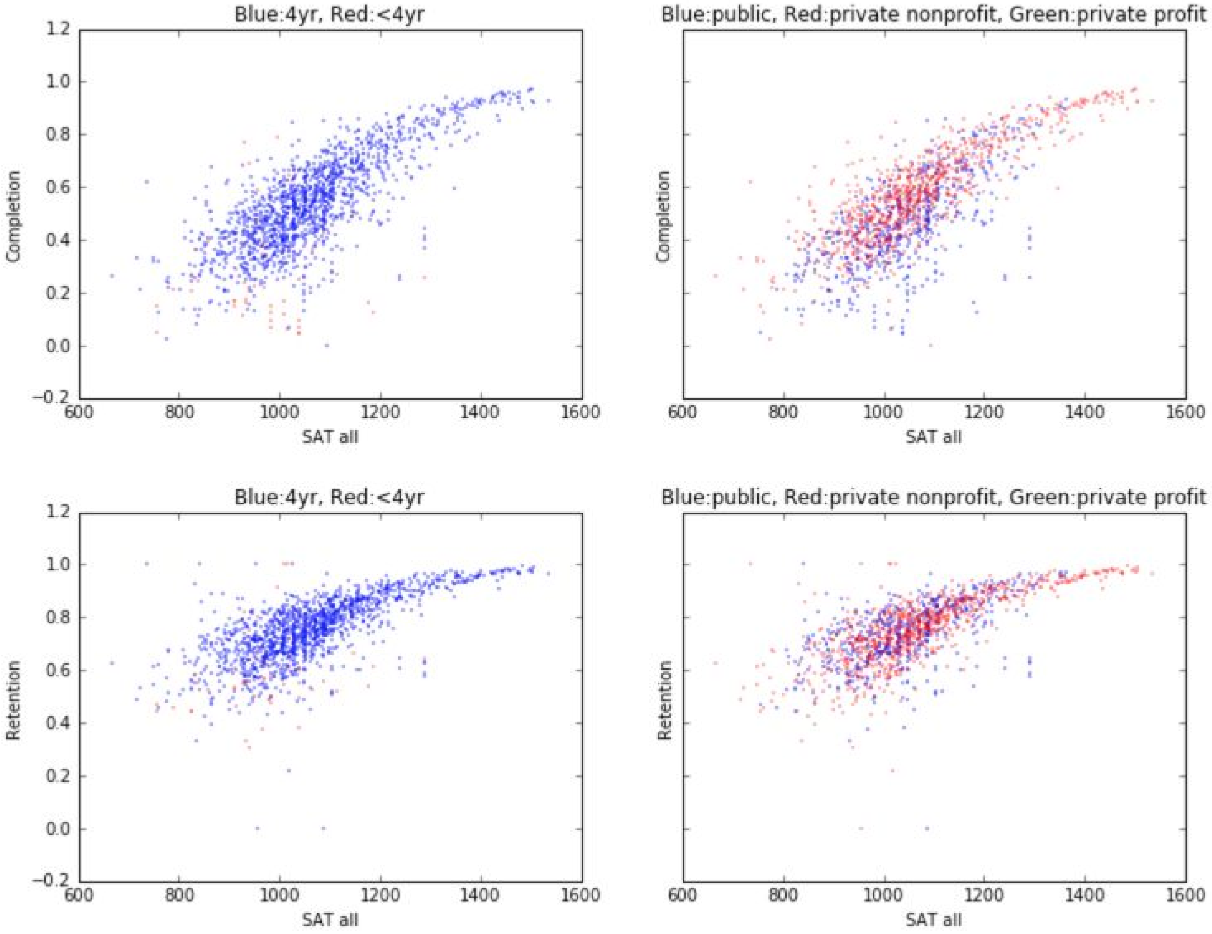


There is a pretty good correlation between average faculty salary and retention especially when we look at it for each different college type. For example for <4 year college, higher faculty salary actually correlate with lower completion rate, but for 4 year college it's the opposite.

SAT

```
count    1445.000000
mean     1061.523183
std       131.050271
min       666.000000
25%       977.000000
50%      1048.000000
75%      1123.000000
max      1534.000000
```

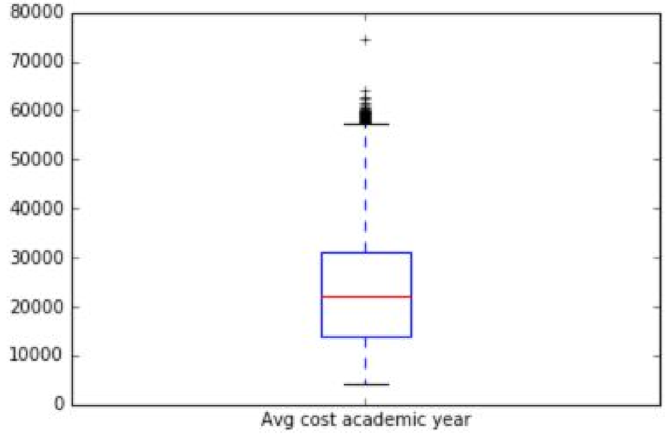


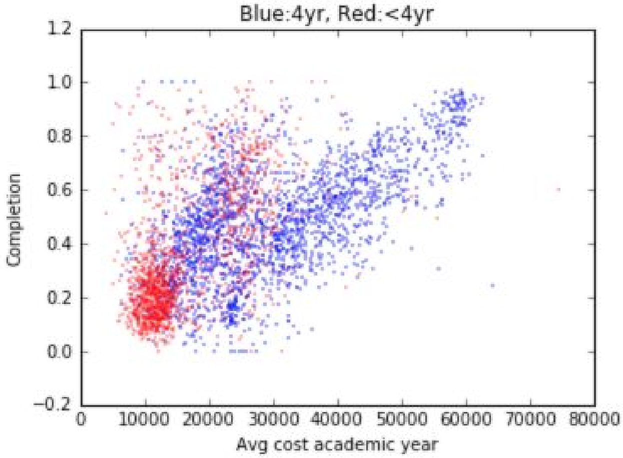


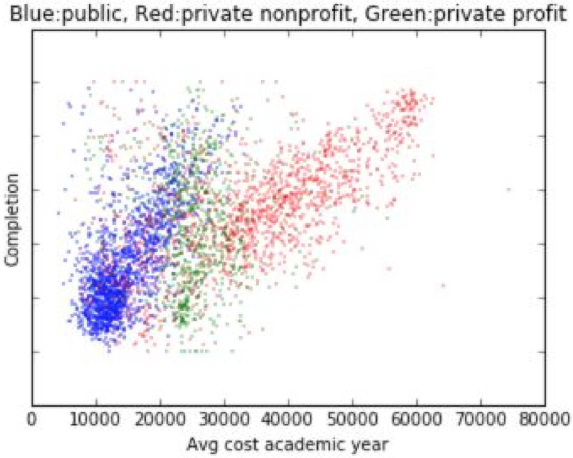
There is a nice correlation between SAT score with the completion rate. ACT score has similar pattern (see exploratory.ipynb).

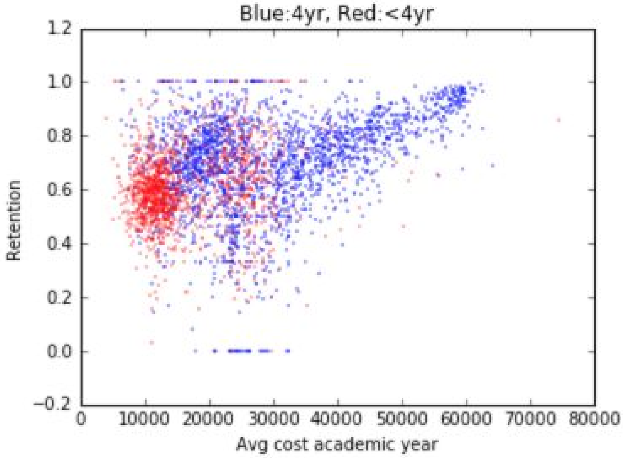
Average cost academic year

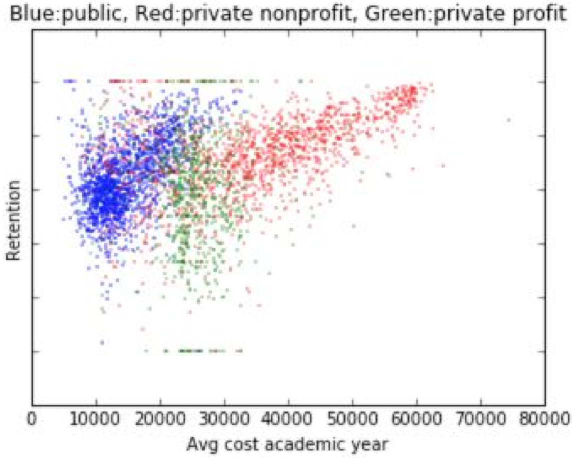
```
count      3691.000000
mean      24397.743430
std       12682.727804
min        4157.000000
25%       13734.000000
50%       22188.000000
75%       31251.000000
max       74473.000000
```




Blue:4yr, Red:<4yr


Blue:public, Red:private nonprofit, Green:private profit

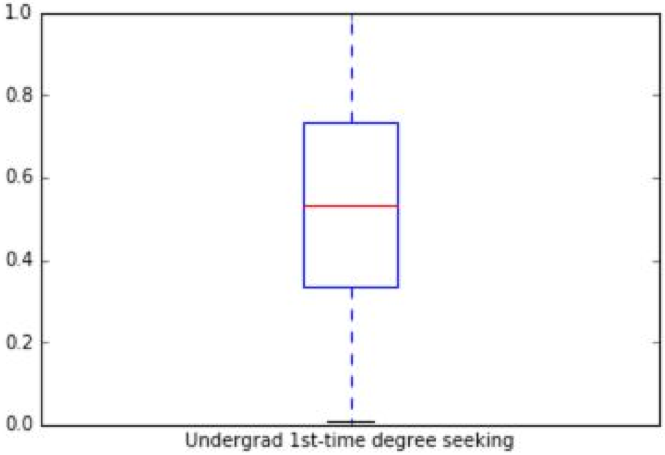

Blue:4yr, Red:<4yr

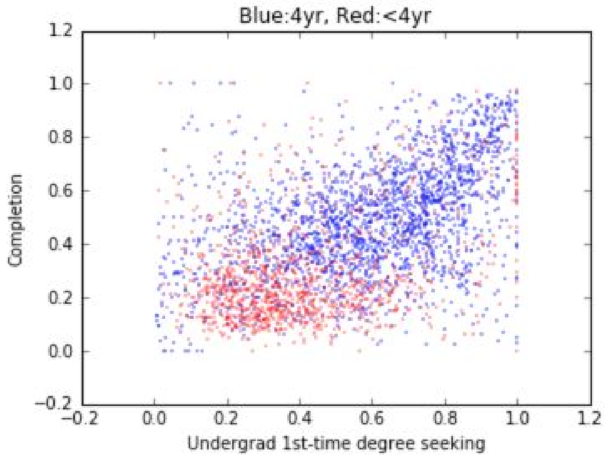

Blue:public, Red:private nonprofit, Green:private profit

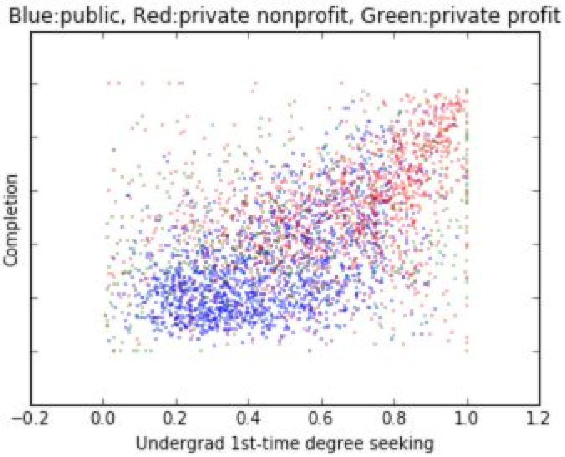
There is an interesting correlation for Average cost academic year. Looking at the right chart, it looks like each category (public, private, profit) has its own correlation for cost vs completion/retention rate.

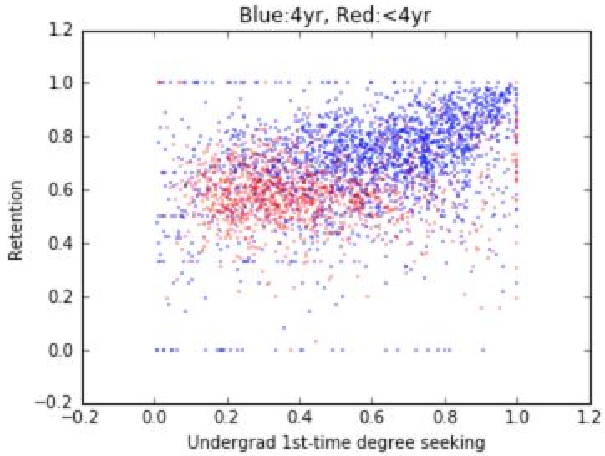
Undergrad 1st time degree seeking

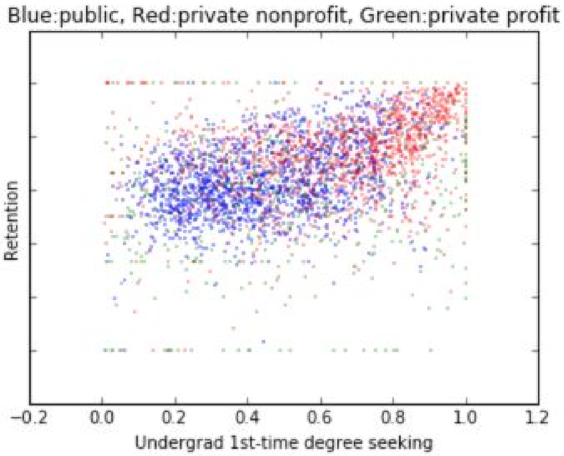
```
count    3290.000000
mean        0.534137
std         0.245144
min         0.008600
25%         0.336250
50%         0.533950
75%         0.736025
max         1.000000
```







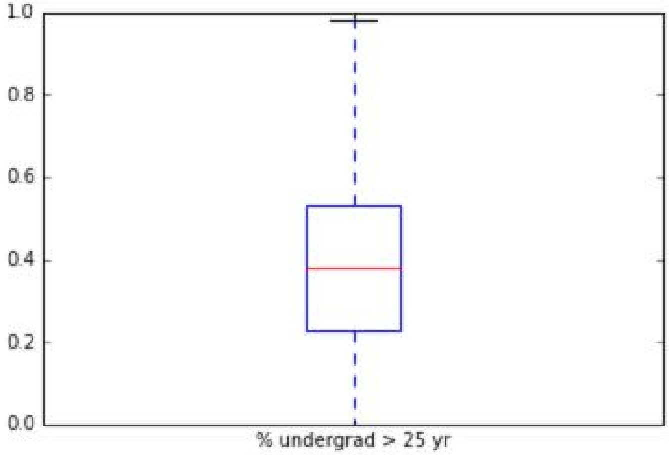


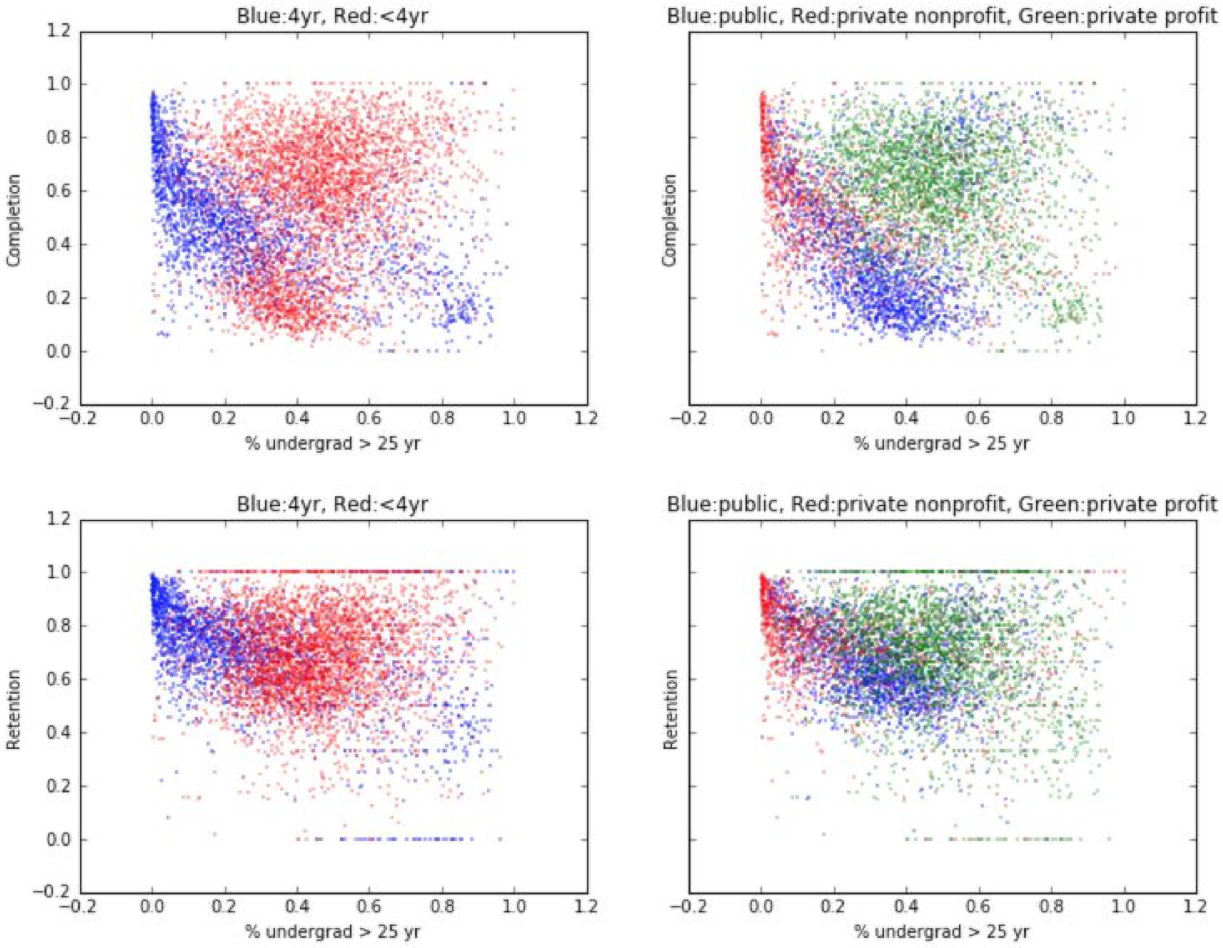


There is a pretty good correlation for Undergrad 1st time degree seeking.

Percentage of undergrad > 25 year old

```
count    5964.000000
mean        0.386220
std         0.217355
min         0.000500
25%         0.228600
50%         0.379100
75%         0.532950
max         1.000000
```




Blue:4yr, Red:<4yr

Blue:public, Red:private nonprofit, Green:private profit

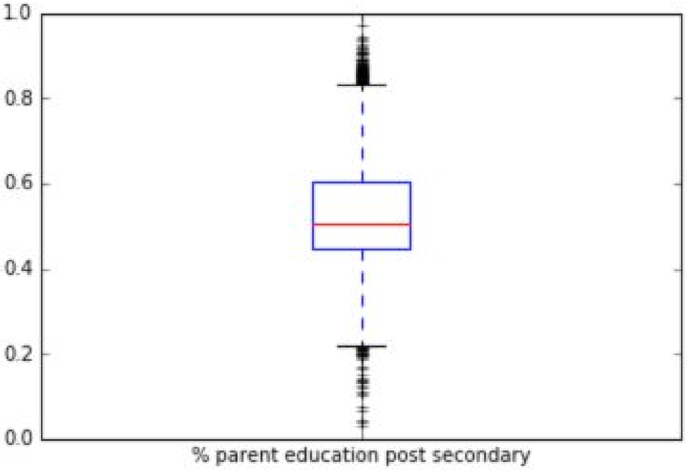
Blue:4yr, Red:<4yr

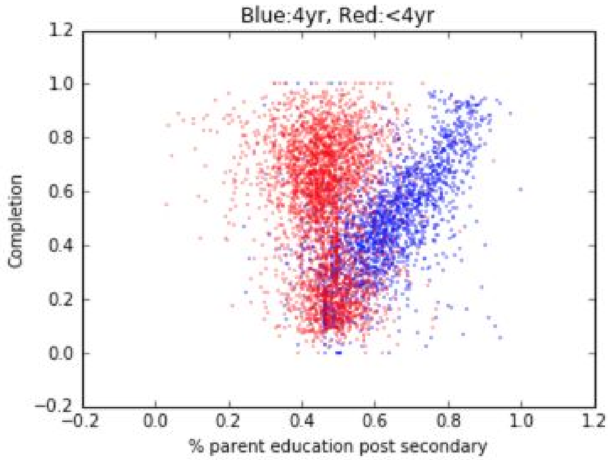
Blue:public, Red:private nonprofit, Green:private profit

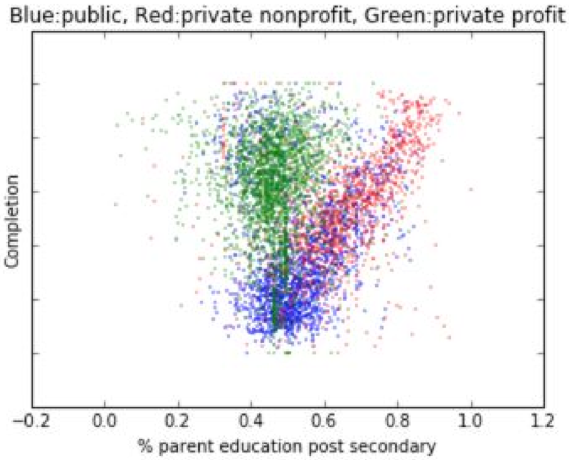
Four year colleges, public and private colleges have a strong negative correlation between Percentage of undergrad > 25 year. However “for profit college” does not have the same correlation. It has more percentage of > 25 year but higher completion/retention.

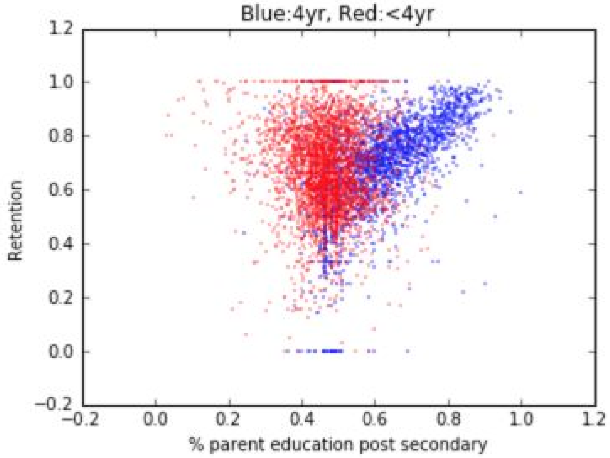
Percentage of parent education post secondary

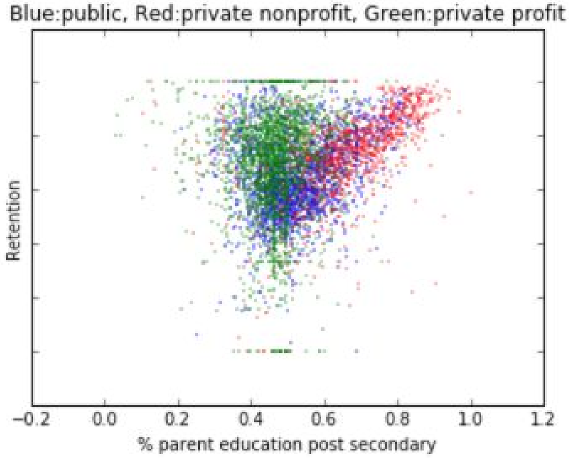
```
count    5533.000000
mean        0.529217
std         0.128194
min         0.030588
25%         0.447724
50%         0.504145
75%         0.602070
max         1.000000
```











There is a good correlation between Percentage of parent education post secondary and completion rate.

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. It also performs well with large data set.
- 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. One disadvantage of SVM is that it is computationally expensive for a large data set.
- K-Nearest Neighbor
  This algorithm should do pretty well in predicting the retention and completion rate based on how other similar colleges do. One downside is that KNN is slow on prediction time when we have a large data set.
- Random Forest
  Random Forest has most of the benefits from Decision Tree: scale well for big data set, does not have much parameters to tune. In most cases it also has better performance than Decision Tree.

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

Benchmark

We will use the Median Absolute Error (MAE) 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

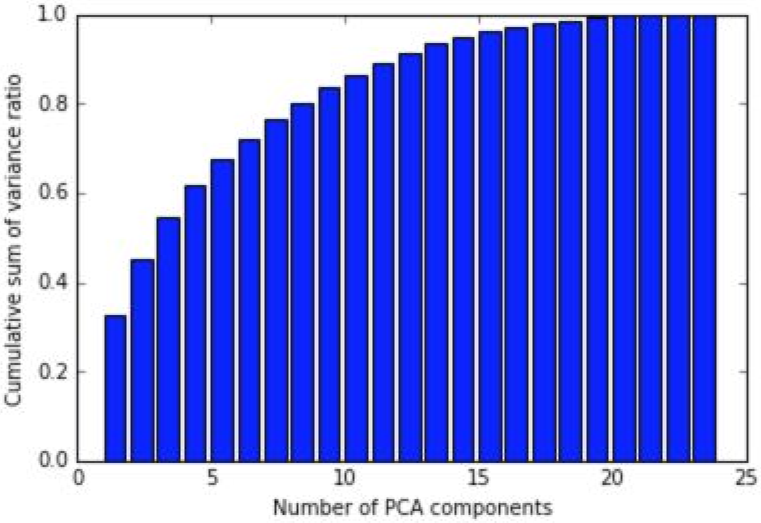
Because there are a lot of missing values, we'll fill the missing values with the mean of the feature. For example, there are only about 1500 colleges with SAT score, so we'll fill the missing SAT scores with the average of the SAT score from the 1500 colleges. This method for imputing missing data have the possibility of obscuring the correlation in the data, especially for features that have relatively large number of missing data.

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 zero 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. Note that the first 19 PCA components explains almost 100% of the variance, but using 19 PCA components does not give better result. In one case, using the same train/test data, 19 PCA components gives slightly worse result than 15 PCA components.

One hot encoded categorical features were not included in the PCA in this first model, because at first my intuition says that it would cause some important categorical features to be lost when we reduce them with PCA. (See the refinement section where one hot encoded features are also included in the PCA)

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)
```

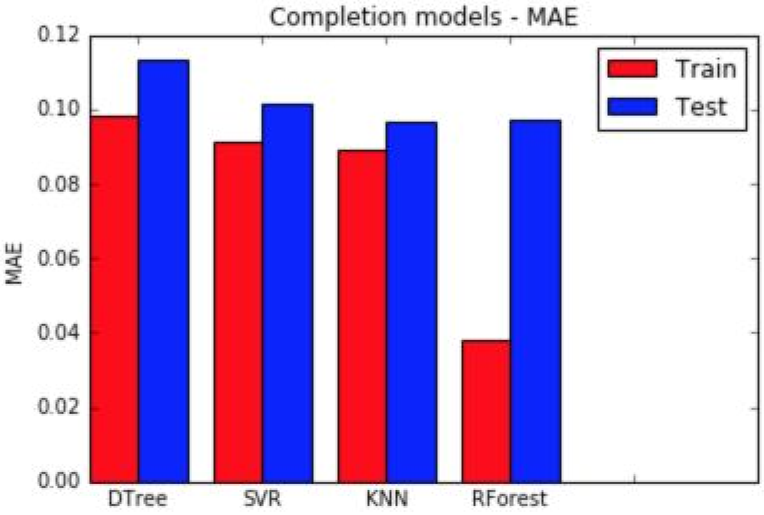
The y_train and y_test are Nx2 arrays containing the completion and retention data. This helper method also accept number of cross validation and params for doing Grid Search. Each of this helper function will return an object that has the regression models, R2, MSE and MAE scores for each model (for both test and train data).

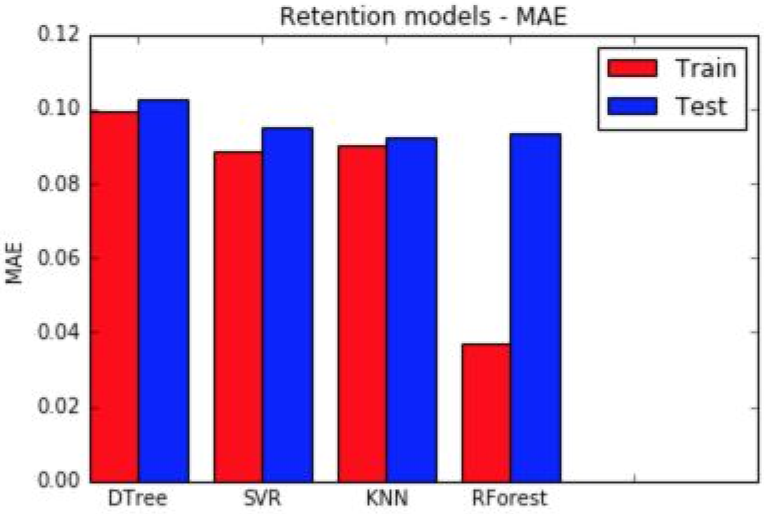
The preprocessed data consist of:

- 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 result for SVR model:

Completion MAE: 0.102
Retention MAE: 0.094



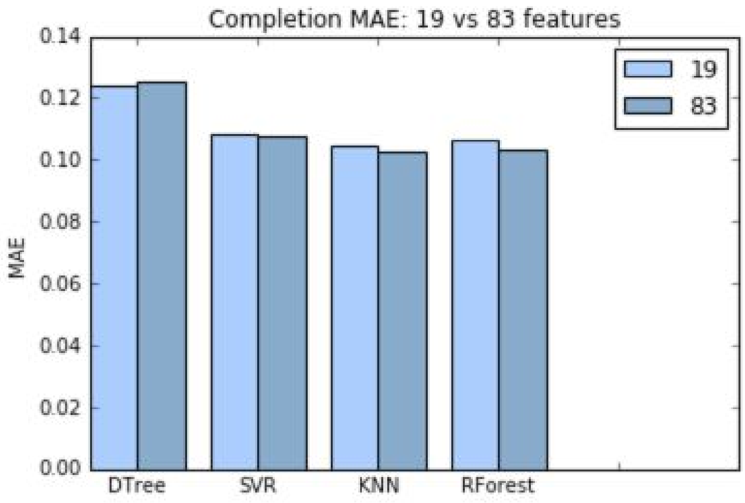


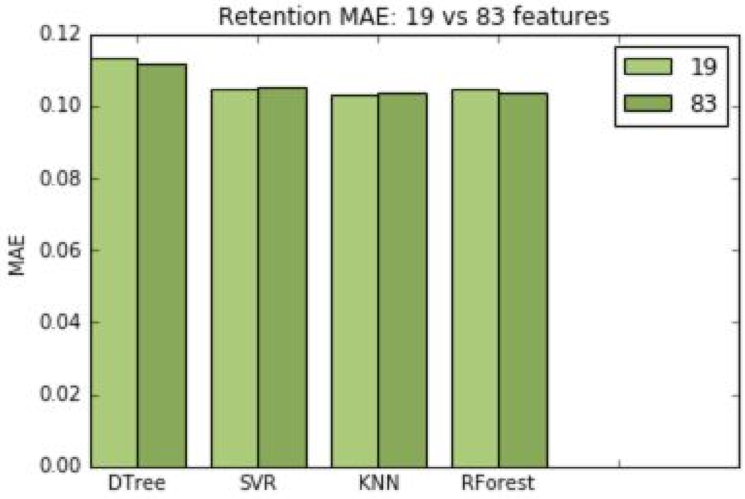
Depending on train/test randomization, SVR, KNN and Random Forest's result can be different (there is no clear winner). However, generally they all have very close MAE score (around 0.1 for Completion model and slightly less than 0.1 for Retention model)

Refinement

Adding More Features

The first refinement is to add more features. All 3 Carnegie classifications are added (as one hot encoded) which result in 83 features total. This new input is fed to the same helper methods that build the models. The result of adding more features is almost negligible. There are cases (depending on the train/test randomization) where more features perform slightly worse. Here is one performance comparison between models with 19 and 83 features:





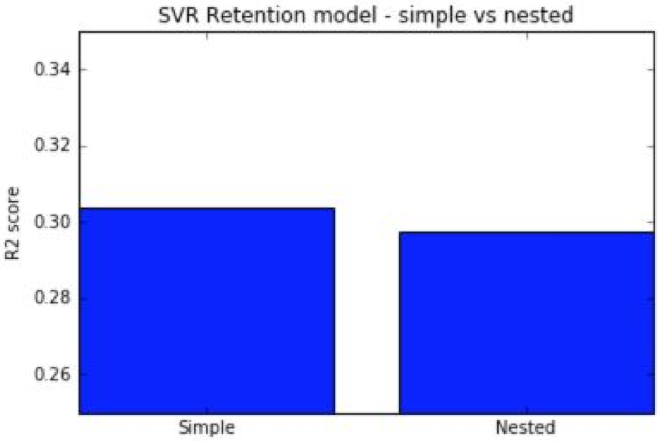
Using Completion model to help predict Retention

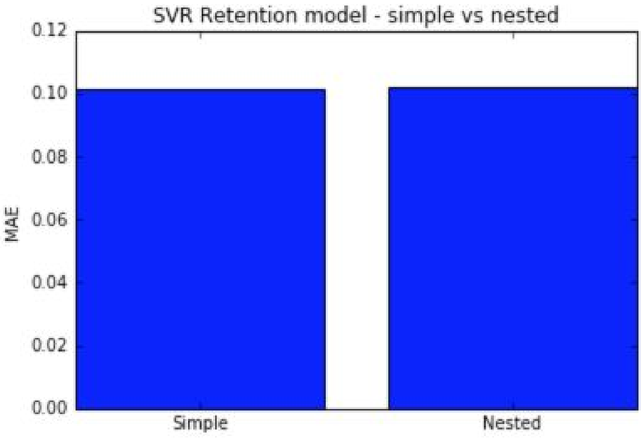
Even though the MAE for Completion and Retention are about the same, the Completion model has a much better R2 score than the Retention model. We could use the Completion model to help predict the Retention. This might work because there is a clear positive linear correlation between completion and retention data. Let's call this "SVR nested model" because we use the Completion model to help predict the 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.

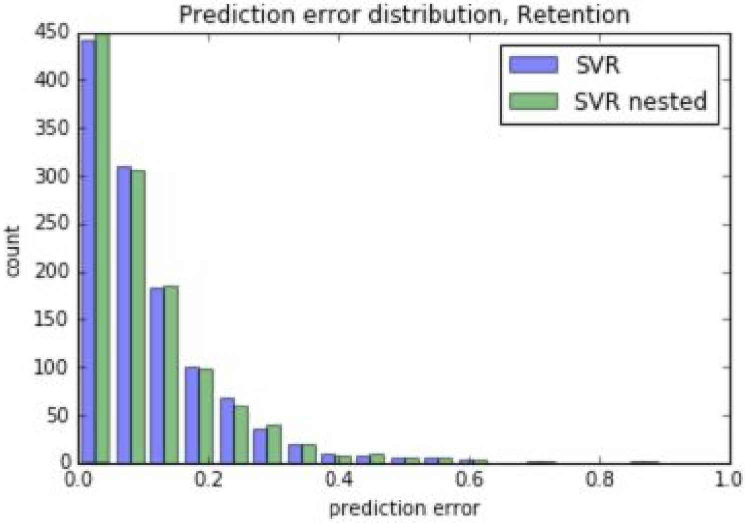
SVR Nested Result

The result is not very promising, in some cases the R2 score improve a little but sometimes it also become worse. The MAE is mostly about the same.





Looking at Retention prediction error distribution, the nested SVR model have roughly the same error distribution as the simple SVR.



Better Feature Selection / Building Model From Reduced Dataset

Because one of the challenge in this problem is a lot of missing values, it's worth trying how good we can get by reducing the data to only those that have complete features. In order to do this we only use even less features that are selected based on a strong correlation from the data visualization earlier.

In `model-reduced-dataset-2.ipynb`, only 12 features are selected. Two of the features are categorical: college type (public, private, profit) and whether it is 4 year or less than 4 year college. After removing data that have missing values for any of the 12 features, 1242 rows remain. We then one hot encoded the categorical feature, so the final data is a 1242 x 14 matrix.

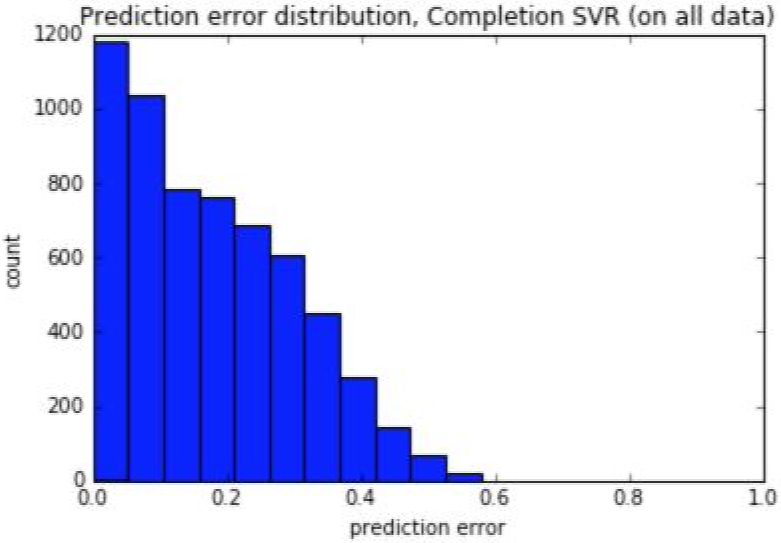
The same standards scaler is used, but we do not apply PCA for dimensional reduction. Then we split the data into train / test with 0.8 ratio. The same `build_SVR_model` helper method is used to build SVR regression model using this smaller dataset.

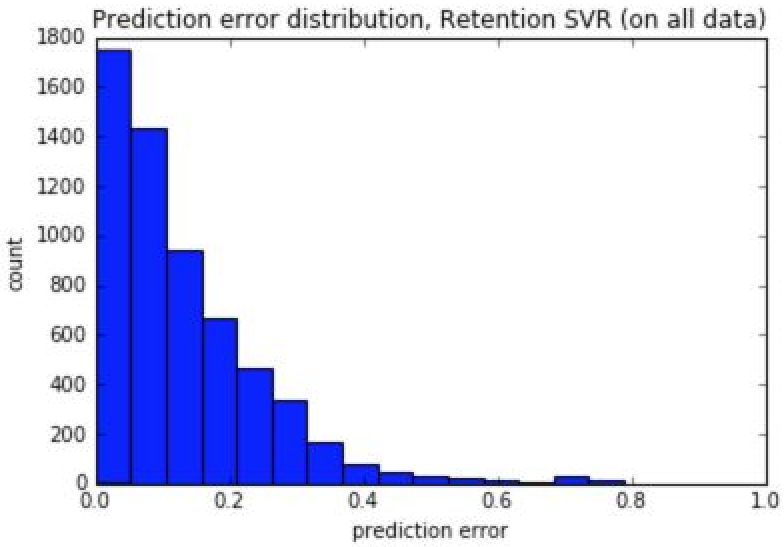
Feature Selection Result

The result is a much better model with MAE of 0.06 and 0.05 for Completion and Retention model respectively when tested on the test set from the reduced data. This is much better than MAE of 0.1 for our best model. This is not a surprise because it only makes prediction on limited data with no missing values. So we will try using the model trained on the smaller dataset to predict larger dataset that has missing values. First we will fill the missing values on the larger dataset with the mean. The result is:

Completion MAE: 0.18
Retention MAE: 0.13

This is worse than MAE of 0.1 from our previous best model. Let's analyze how this model perform much worse on the larger dataset.





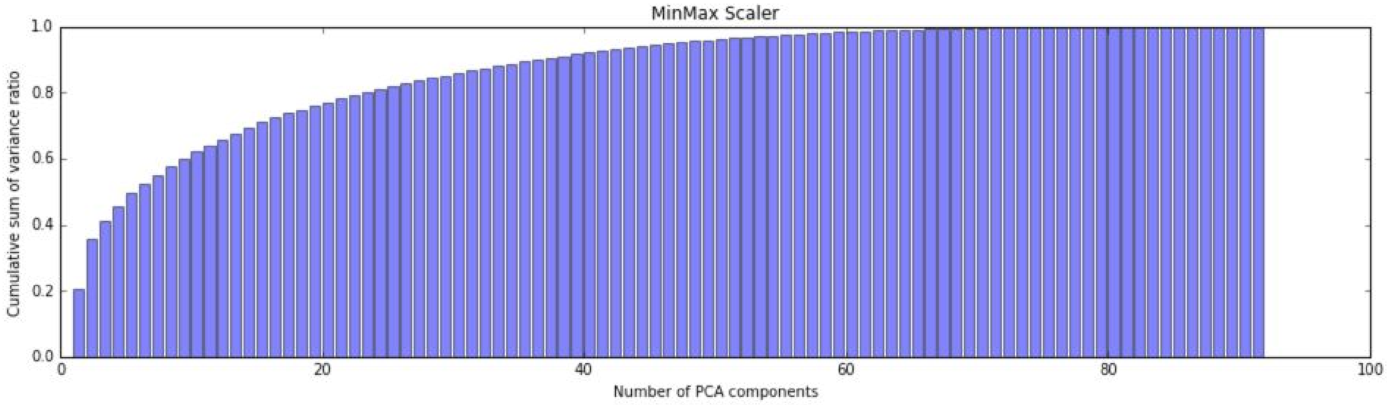
The first chart (Completion model) shows that the model actually does quite well for a number of data with 2200 case having less than 0.1 prediction error. But there are still a large number of data with prediction error > 0.2. The Retention model is a little bit better but also suffers from prediction error > 0.2 for quite a large number of data.

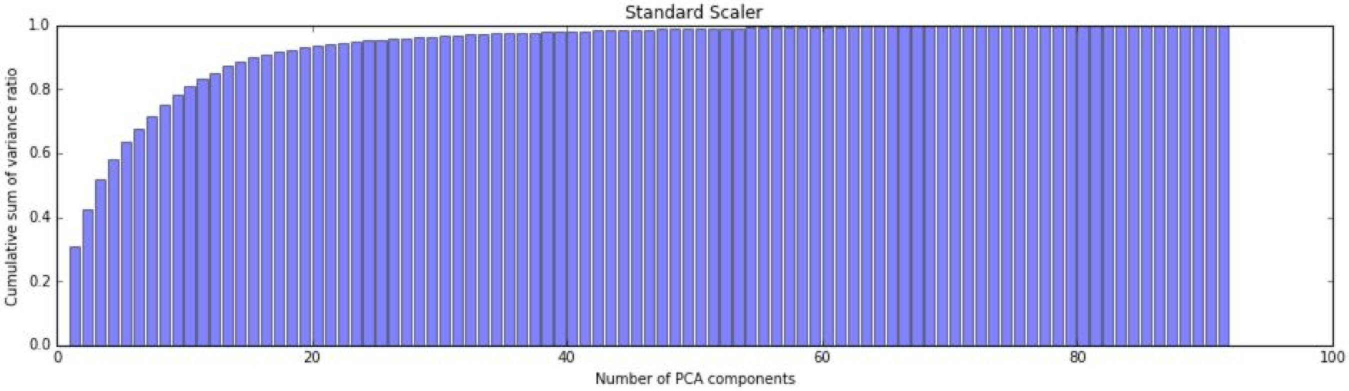
Including One Hot Encoded Features in PCA

As suggested by Udacity coach from the review, I tried including one hot encoded features in PCA. I also tried 2 different approach:

- Use Standard scaler for numerical features, the same as previous model (model-final-onehotPCA-StandardScaler.ipynb)
- Use Min Max scaler for numerical features (model-final-onehotPCA-MinMaxScaler.ipynb)

There are a total of 91 numerical and categorical features that are included in the PCA. Here is the visualization of the cumulative sum of variance ratio of the PCA components for each approach:





The cumulative sum of variance ratio of the PCA components for Standard Scaler approaches 100% faster than MinMax Scaler. This means that we can use fewer PCA components when using Standard Scaler. So using Standard Scaler, we choose the first 35 PCA components to build the model because they account for more than 95% of the variance.

The result is about the same with the original model. MAE score for SVR model:

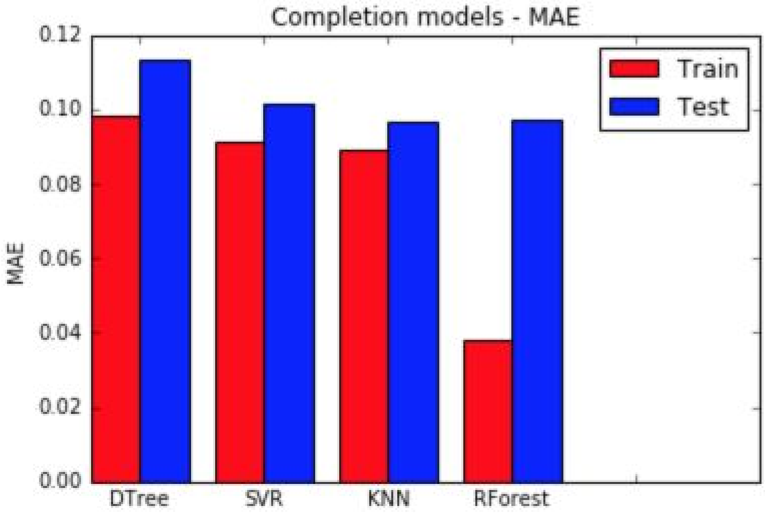
Completion MAE: 0.101
Retention MAE: 0.096

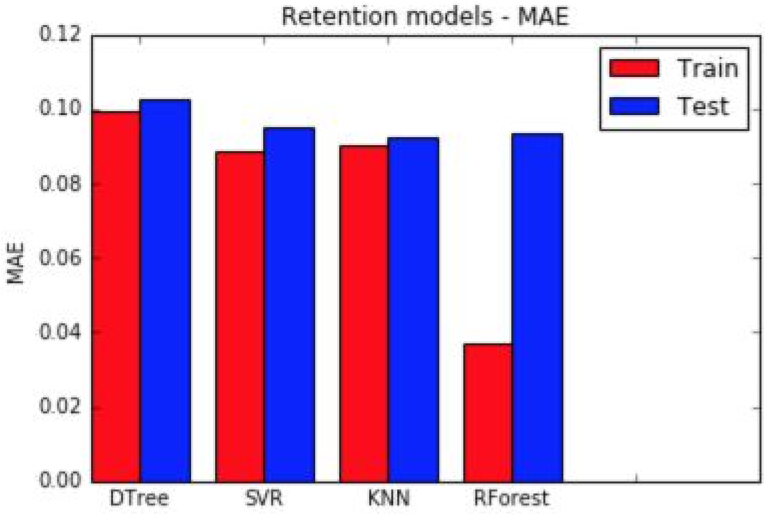
Results

Model Evaluation and Validation

Because the refinement attempt does not produce model that performs noticeably better than the original model for predicting the larger dataset, we will focus on analyzing the original models.

Even though the three models (SVR, KNN and RandomForest) have similar MAE 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 MAE scores. SVN and KNN have training and testing MAE scores that converge better.

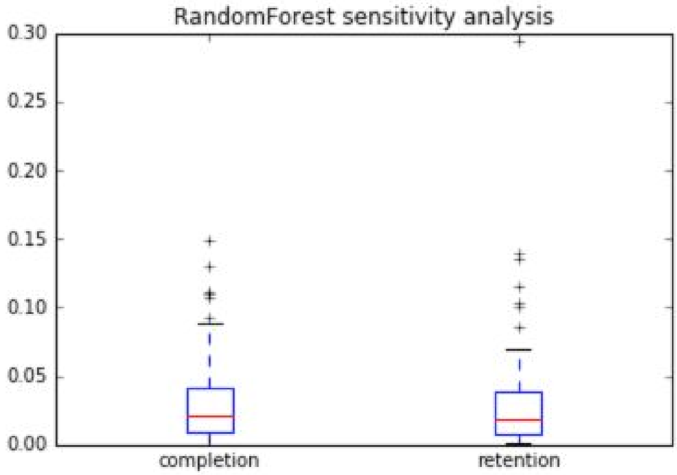


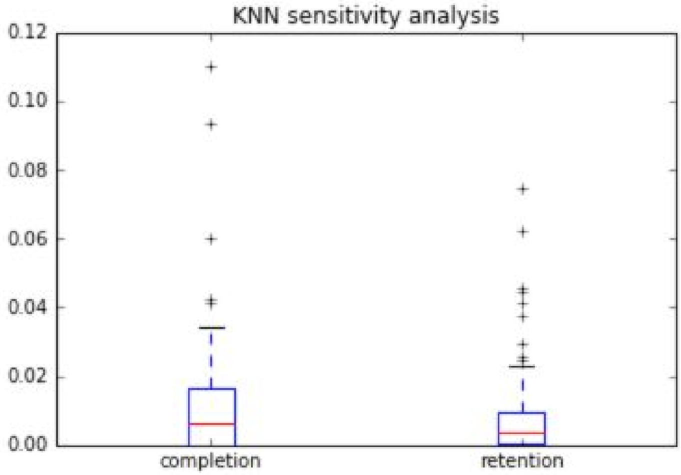


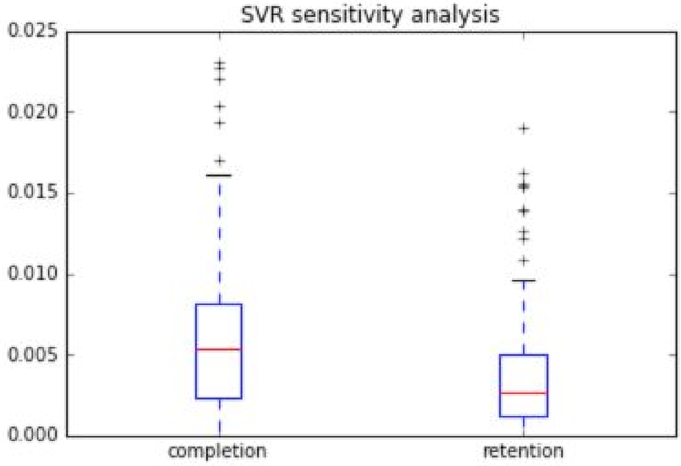
Sensitivity Analysis

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.







Final Model

The final model chosen is the SVR that uses 19 features with the following parameters (for both completion and retention):

- epsilon = 0.1
- C = 0.1
- gamma = 0.1

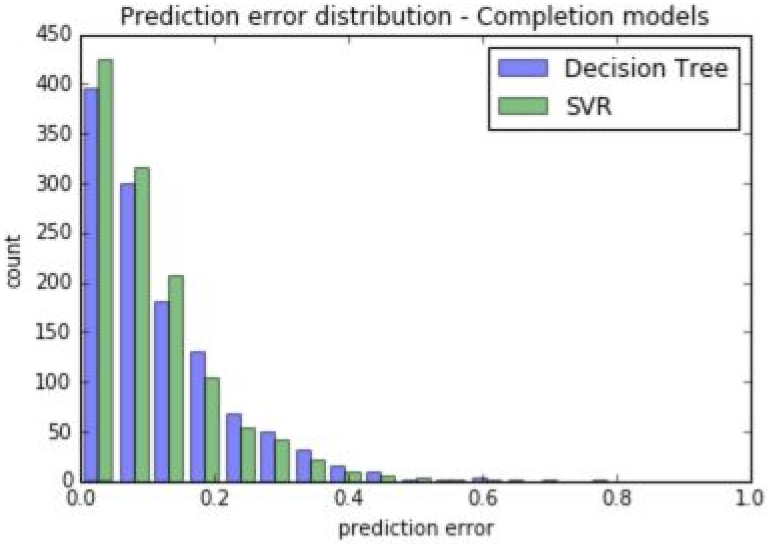
A few reasons SVR was chosen:

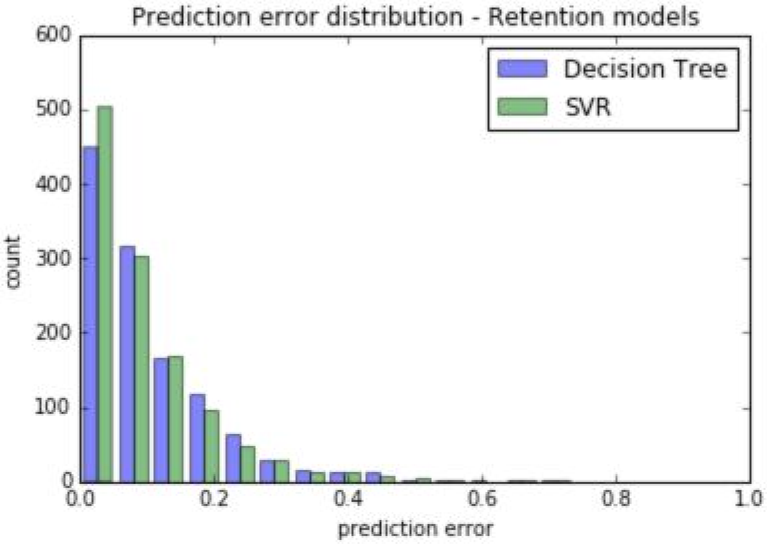
- It consistently performs well enough. It does not always perform the best (when compared with KNN and Random Forest), but it never performs much worse than KNN and Random Forest either.
- The MAE score between train and test data converges pretty well for both completion and retention model.
- SVR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

Justification

The final SVR model has MAE scores of about 0.1 for Completion and about 0.095 for Retention model. That is a slight improvement from the benchmark (simple decision tree model) MAE scores of 0.12 and 0.1 for Completion and Retention respectively.

Comparison of prediction error distribution between the benchmark model (Decision Tree) with the final SVR Model: (this is done on the 1192 test data)





Looking at prediction error distribution

- For Completion model about 61% of the model prediction error is < 0.1 and 25% between 0.1 and 0.2.
- For Retention model about 67% of the prediction error is < 0.1 and 23% between 0.1 and 0.2

Considering completion and retention rate of a college as something that fluctuates from year to year, probably by as much as 0.1 or more for colleges with very few students. We can conclude that our model performance with Median Absolute Error of 0.1 adequately addresses the problem. It's good enough to get an idea if the school has a high, medium or low retention and completion rate.

Conclusion

Reflection

The first step on addressing this problem is to pick a subset of features from 1700 available features in the dataset by reading the data dictionary and deciding which features might influence the Completion and Retention rate of a college. After selecting a subset of features, we visualize the correlation of each feature with the target features and further filter the features for building the models.

First attempt at building the model just uses the selected numerical features (reduced to 15 features using PCA) and 2 categorical features (which are one hot encoded resulting in 4 features). This turned out to give a good enough result that it become the selected final model.

A few refinements attempt were made:

1. Adding more categorical features. There are 3 more categorical features which adds 62 features when one hot encoded. The results sometimes show a slight improvement but there are cases where it performs worse than the original model with much fewer features.
2. Using the Completion model to build the Retention model. This does not give a noticeable improvement to the Retention model. In some cases it actually causes a worse R2 score even when the MAE stays the same.
3. Build models using a reduced dataset where there is no missing values. This gives a much better result on the smaller dataset because it eliminates the problem we have when imputing the missing data by filling with the means. However this model performs much worse than the original model on the larger dataset.
4. Including one hot encoded features in the PCA. This generally has similar performance to the original model.

A few algorithms were used during this process: Decision Tree, SVR, KNN, and Random Forest. SVR was selected not because it always performs better, but because of a good trade off between performance (MAE score), stability on sensitivity analysis as well as prediction speed (faster than KNN).

One interesting aspect in this project is that using R2 score to judge performance could be misleading. In all cases the R2 score for Retention is much lower than that of Completion model. However the MAE score for Retention model is generally better than the Completion model. This is caused by the distribution of the Retention data, which is more concentrated in the 0.3 - 1 range with some outliers below 0.3. The distribution of the Completion data is more evenly spread out between 0 and 1.

One difficult aspect 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-2.ipynb` I was able to reduce the dataset to 1242 rows that have all the data for a subset of the features. The SVR models from this reduced dataset actually get much better MAE scores for both completion and retention prediction (0.062 and 0.055 respectively). However this model fails to perform well on the larger dataset.

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 MAE 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 MAE scores between the new baseline benchmark and the target benchmark.