Predicting fraction of students in 4-year colleges who complete the program in less than 6 years

Abhinav Singh

December 15, 2016

Motivation

My goal is to build a model that would accurately predict the fraction of students enrolled in a 4-year colleges and complete the program in much less than 6 years based on the specific features of the college like rate of admission for all demographics (ADM_RATE_ALL in the DB), total cost of attendance per year average, acceptance rate, average test scores etc. Because of the increasing cost of higher education and the accompanying rise of student debt, it is important to understand the factors that contribute to the post-graduation rate and ability of the students and delve into their ability to pay education loans. This model will increase the transparency and will help students to weigh the trade-offs of different colleges and make more informed decisions.

Background

I am using the College Scorecard Data [1] originally curated by US Department of Education. This dataset which spans nearly 20 years, contains approximately 2000 features for 7805 degree-granting institutions. These features include different information like demographic data, test scores, family income data, and data about the percentages of students in each major, financial aid information, debt and debt repayment values, earnings of alumni several years after graduation.

Data pre-processing

The dataset contains information of about 8000 colleges, and for each college there are around 2000 features. But the data is also sparse, containing a large number of missing and privacy suppressed values. So we filtered out the columns whose fraction of bad (missing or privacy suppressed) values is greater than a certain threshold. Even though we have data from 1996 to 2013, I am using data from year 2011(training) and 2013(testing) as these include lot required variables for analysis.

Discovering significant features

Since we have a huge number of features, it is unlikely that all of them would affect the post-graduation completion. So to see which are the most significant features for predicting the post-graduation completion, I did feature selection. I used forward stepwise selection to select top features.

I also plan on using multivariate machine learning techniques, which is also referred to as MVA. I want to predict a fraction, so this is a regression problem. I use college and aggregated student data from 2011 to train 3 different MVA methods:

- single decision tree
- random forest of decision trees
- support vector machine (SVM)

The learning is subsequently evaluated on 2013 data. Below are details of the workflow.

The variable that I want to predict (variable of interest VOI) is **C150_4** in the 2011 dataset. In the first step, I look at all quantitative variables associated with colleges and (aggregated) students. 14 variables look promising in terms of predictive power as well as per PCA:

1. rate of admission for all demographics (ADM_RATE_ALL)
2. total cost of attendance per year (COSTT4_A)
3. fraction of students that are white (UGDS_WHITE)
4. fraction of students that are black (UGDS_BLACK)
5. fraction of part-time students (PPTUG_EF)
6. fraction of students from low-income families, defined as annual family income < $30,000 (INC_PCT_LO)
7. fraction of students above 25 years of age (UG25abv)
8. fraction of first-generation college goers in the family (PAR_ED_PCT_1STGEN)
9. fraction of students on federal loan (PCTFLOAN)
10. 75% percentile score on SAT reading (SATVR75)
11. 75% percentile score on SAT writing (SATWR75)
12. 75% percentile score on SAT math (SATMT75)
13. tuition fee per year for in-state students (TUITIONFEE_IN)
14. tuition fee per year for out-of-state students (TUITIONFEE_OUT)

Dataset Description

I downloaded the data from [1] and there are around 18 csv files starting from year 1996 to 2013. Here are certain snippets from the dataset:

UNITID	OPEID	opeid6	INSTNM	CITY	STABBR	ZIP	AccredAge	INSTURL	NPCURL	sch_deg	HCM2	main	NUMBRAN	PREDDEG	HIGHDEG	CONTROL	st_fips	region	LOCALE	locale2	LATITUDE	LONGI
100636	1230800	12308	COMMUN	MONTGOI	AL	3.61E+08	NULL	NULL	NULL	NULL	0	1	1	2	2	1	1	0	NULL	NULL	NULL	NULL
100654	100200	1002	ALABAMA	NORMAL	AL	35762	NULL	NULL	NULL	NULL	0	1	1	3	4	1	1	5	NULL	NULL	NULL	NULL
100663	105200	1052	UNIVERSI	BIRMINGH	AL	3.53E+08	NULL	NULL	NULL	NULL	0	1	2	3	4	1	1	5	NULL	NULL	NULL	NULL
100672	574900	5749	ALABAMA	OZARK	AL	36360	NULL	NULL	NULL	NULL	0	1	1	1	2	1	1	5	NULL	NULL	NULL	NULL
100690	2503400	25034	SOUTHERI	MONTGOI	AL	3.61E+08	NULL	NULL	NULL	NULL	0	1	1	3	4	2	1	5	NULL	NULL	NULL	NULL
100706	105500	1055	UNIVERSI	HUNTSVIL	AL	35899	NULL	NULL	NULL	NULL	0	1	1	3	4	1	1	5	NULL	NULL	NULL	NULL
100724	100500	1005	ALABAMA	MONTGOI	AL	3.61E+08	NULL	NULL	NULL	NULL	0	1	1	3	4	1	1	5	NULL	NULL	NULL	NULL
100751	105100	1051	THE UNIVI	TUSCALOC	AL	3.55E+08	NULL	NULL	NULL	NULL	0	1	1	3	4	1	1	5	NULL	NULL	NULL	NULL
100760	100700	1007	CENTRAL /	ALEXANDE	AL	35010	NULL	NULL	NULL	NULL	0	1	1	1	2	1	1	5	NULL	NULL	NULL	NULL
100812	100800	1008	ATHENS S	ATHENS	AL	35611	NULL	NULL	NULL	NULL	0	1	1	3	3	1	1	5	NULL	NULL	NULL	NULL
100830	831000	8310	AUBURN U	MONTGOI	AL	3.61E+08	NULL	NULL	NULL	NULL	0	1	1	3	4	1	1	5	NULL	NULL	NULL	NULL
100858	100900	1009	AUBURN U	AUBURN U	AL	36849	NULL	NULL	NULL	NULL	0	1	1	3	4	1	1	5	NULL	NULL	NULL	NULL
100919	570400	5704	BESSEMER	BESSEMER	AL	35021	NULL	NULL	NULL	NULL	0	1	1	1	2	1	1	5	NULL	NULL	NULL	NULL
100937	101200	1012	BIRMINGH	BIRMINGH	AL	35254	NULL	NULL	NULL	NULL	0	1	1	3	4	2	1	5	NULL	NULL	NULL	NULL
100964	573301	5733	BEVILL ST	FAYETTE	AL	35555	NULL	NULL	NULL	NULL	0	0	3	3	4	1	1	5	NULL	NULL	NULL	NULL
101019	2319500	23195	CHARLES A	MOBILE	AL	36606	NULL	NULL	NULL	NULL	0	1	1	1	1	3	1	5	NULL	NULL	NULL	NULL
101028	1218200	12182	CHATTAHC	PHENIX CI	AL	36869	NULL	NULL	NULL	NULL	0	1	1	2	2	1	1	5	NULL	NULL	NULL	NULL
101037	573600	5736	CHAUNCE	EUFAULA	AL	36027	NULL	NULL	NULL	NULL	0	1	1	1	2	1	1	5	NULL	NULL	NULL	NULL
101073	1055400	10554	CONCORD	SELMA	AL	36701	NULL	NULL	NULL	NULL	0	1	1	2	3	2	1	5	NULL	NULL	NULL	NULL
101107	569800	5698	DOUGLAS	OPP	AL	36467	NULL	NULL	NULL	NULL	0	1	1	1	2	1	1	5	NULL	NULL	NULL	NULL
101116	446300	4463	DRAUGHO	MONTGOI	AL	36104	NULL	NULL	NULL	NULL	0	1	1	2	2	3	1	5	NULL	NULL	NULL	NULL
101143	101500	1015	ENTERPRI	ENTERPRI	AL	36330	NULL	NULL	NULL	NULL	0	1	1	2	2	1	1	5	NULL	NULL	NULL	NULL
101161	106000	1060	JAMES H F	BAY MINE	AL	36507	NULL	NULL	NULL	NULL	0	1	1	2	2	1	1	5	NULL	NULL	NULL	NULL
101189	100300	1003	FAULKNER	MONTGOI	AL	3.61E+08	NULL	NULL	NULL	NULL	0	1	5	3	4	2	1	5	NULL	NULL	NULL	NULL
101198	100301	1003	FAULKNER	BIRMINGH	AL	35205	NULL	NULL	NULL	NULL	0	0	5	3	4	2	1	5	NULL	NULL	NULL	NULL
101204	100302	1003	FAULKNER	FLORENCE	AL	35630	NULL	NULL	NULL	NULL	0	0	5	3	4	2	1	5	NULL	NULL	NULL	NULL
101213	100303	1003	FAULKNER	HUNTSVIL	AL	35805	NULL	NULL	NULL	NULL	0	0	5	3	4	2	1	5	NULL	NULL	NULL	NULL
101231	446200	4462	GADSDEN	GADSDEN	AL	35901	NULL	NULL	NULL	NULL	0	1	2	1	1	3	1	5	NULL	NULL	NULL	NULL

I then loaded the data to a dataframe and ran the following SQL query to extract my dataset:

```
SELECT Year, ADM_RATE_ALL, COSTT4_A, UGDS_BLACK, PPTUG_EF, INC_PCT_LO, UG25abv, PAR_ED_PCT_1STGEN, PCTFLOAN, C150_4
       -- CONTROL, SATVR75, SATWR75, SATMT75, TUITIONFEE_IN, TUITIONFEE_OUT, UGDS_WHITE
FROM Scorecard
WHERE Year = 2011
    AND COSTT4_A != 'PrivacySuppressed' AND COSTT4_A IS NOT NULL
    AND ADM_RATE_ALL != 'PrivacySuppressed' AND ADM_RATE_ALL IS NOT NULL
    AND UGDS_BLACK != 'PrivacySuppressed' AND UGDS_BLACK IS NOT NULL
    AND PPTUG_EF != 'PrivacySuppressed' AND PPTUG_EF IS NOT NULL
    AND INC_PCT_LO != 'PrivacySuppressed' AND INC_PCT_LO IS NOT NULL
    AND UG25abv != 'PrivacySuppressed' AND UG25abv IS NOT NULL
    AND PAR_ED_PCT_1STGEN != 'PrivacySuppressed' AND PAR_ED_PCT_1STGEN IS NOT NULL
    AND PCTFLOAN != 'PrivacySuppressed' AND PCTFLOAN IS NOT NULL
    AND C150_4 != 'PrivacySuppressed' AND C150_4 IS NOT NULL
    /*AND CONTROL != 'PrivacySuppressed' AND CONTROL IS NOT NULL
    AND SATVR75 != 'PrivacySuppressed' AND SATVR75 IS NOT NULL
    AND SATWR75 != 'PrivacySuppressed' AND SATWR75 IS NOT NULL
    AND SATMT75 != 'PrivacySuppressed' AND SATMT75 IS NOT NULL
    AND TUITIONFEE_IN != 'PrivacySuppressed' AND TUITIONFEE_IN IS NOT NULL
    AND TUITIONFEE_OUT != 'PrivacySuppressed' AND TUITIONFEE_OUT IS NOT NULL
    AND UGDS_WHITE != 'PrivacySuppressed' AND UGDS_WHITE IS NOT NULL
    AND female != 'PrivacySuppressed' AND female IS NOT NULL
    AND married != 'PrivacySuppressed' AND married IS NOT NULL*/
```

I collect these variables in an R data frame and examine them for availability, integrity and correlations. The SAT scores are privacy-suppressed for the majority of cases, yielding <500 records in 2011. So I decide to leave them out. Tuition fee per year for both in-state and out-of-state students is highly correlated with the total cost of attendance per year, as expected (correlation coefficients > 90%), so the tuition fee variables are excluded. Finally, the fraction of white students has a very small correlation with the VOI, so that this is excluded as well. That leaves me with 8 training variables and 1143 events to train with.

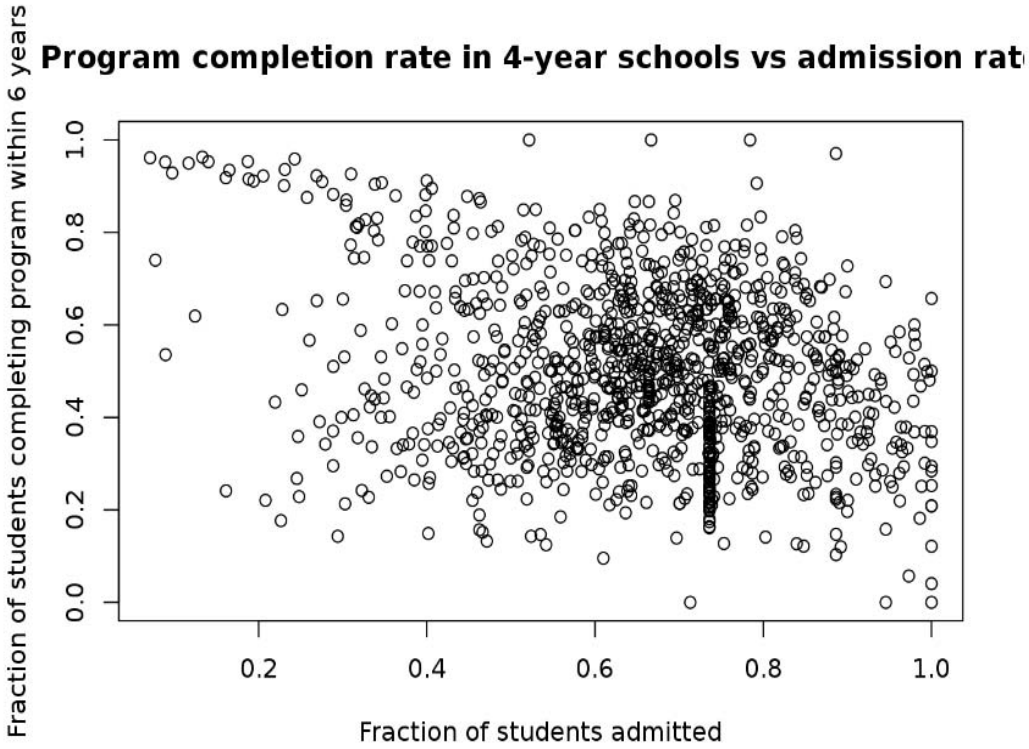
I examine the correlations of these variables with the VOI, and among themselves as necessary. The correlations look like:

Variable	Correlation coefficient with VOI	P-Value
ADM_RATE_AL	0.22	3.77E-15
COSTT4_A	0.57	0.00E+00
UGDS_BLACK	-0.35	0.00E+00
PPTUG_EF	-0.4	0.00E+00
INC_PCT_LO	-0.64	0.00E+00
UG25abv	-0.47	0.00E+00
PAR_ED_PCT_1STGEN	-0.65 (has a 74% correlation with INC_PCT_LO)	0.00E+00
PCTFLOAN	-0.18	9.10E-10

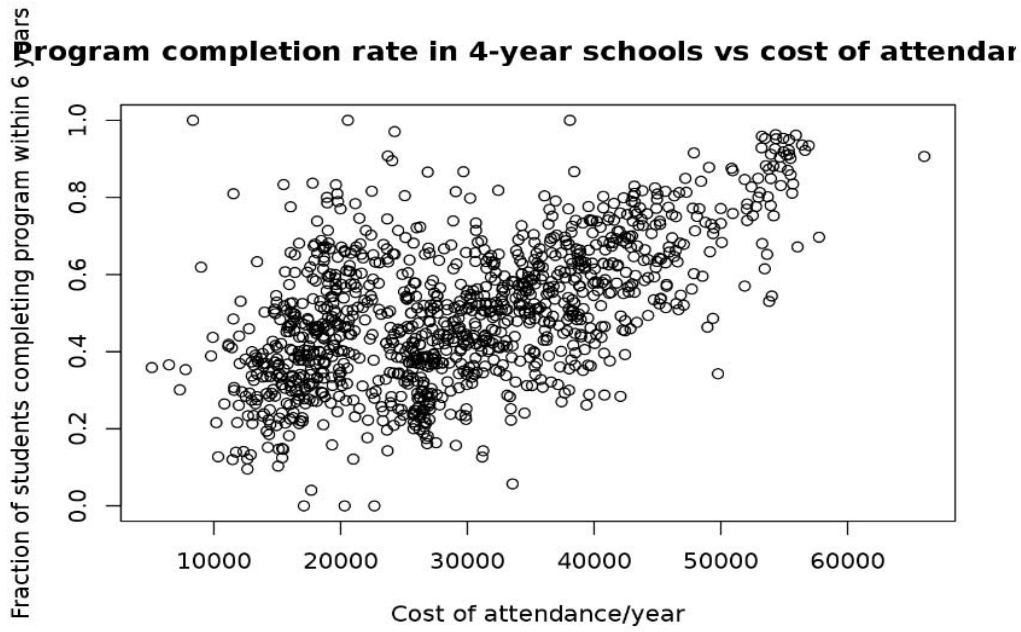
The p-values in the above columns shows that all the variables are significant towards estimating the VOI variable. For some of the variables we see a p-value of 0.000000e+00. Presumably that means the p-value gets rounded down to 0. This happens if the value becomes so small that R cannot represent it anymore using its normal floating point type. Only two of the correlation values are positive and rest else are negative. Negative correlation is a relationship between two variables in which one variable increases as the other decreases, and vice versa. In statistics, a perfect negative correlation is represented by the value -1.00, while a 0.00 indicates no correlation and a +1.00 indicates a perfect positive correlation. E.g.: A simple one would be measuring the amount of snowfall and the temperature. As the temperature increases, the amount of snowfall decreases; this shows a negative correlation and would, by extension, have a negative correlation coefficient. A positive correlation coefficient would be the relationship between temperature and ice cream sales; as temperature increases, so do ice cream sales. This relationship would have a positive correlation coefficient. A relationship with a correlation coefficient of zero, or very close to zero, would be temperature and fast food sales.

Below are some 2D scatter plots for ADM_RATE_ALL, COSTT4_A, INC_PCT_LO and PAR_ED_PCT_1STGEN are shown. This will help us gain more insights in our analysis.

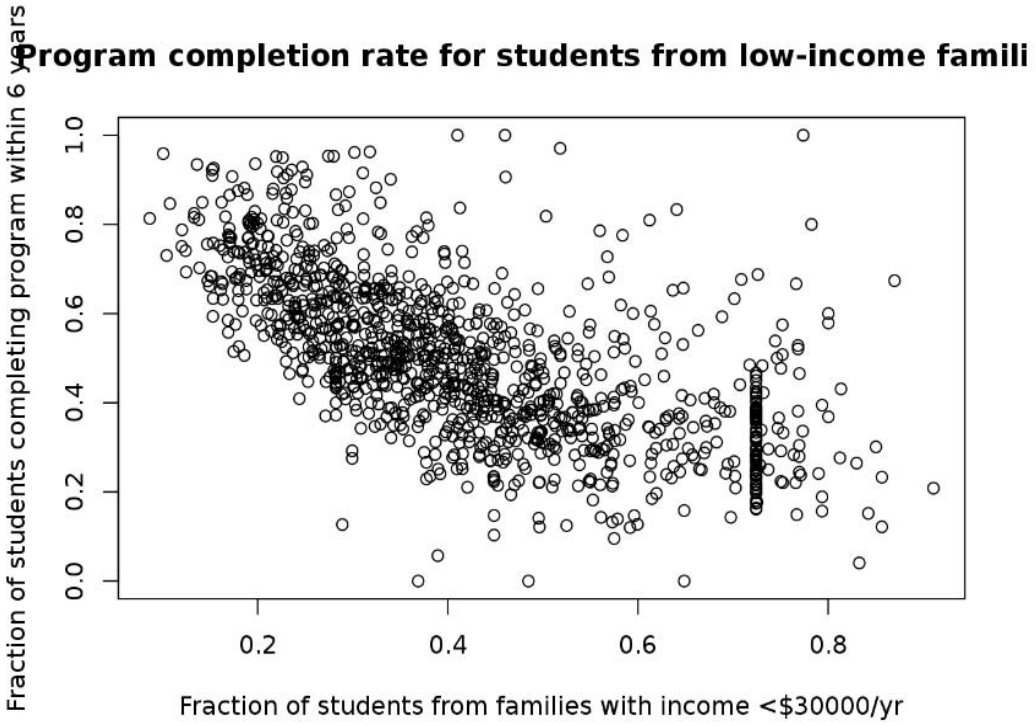
```
plot(train$ADM_RATE_ALL, train$C150_4, main='Program completion rate in 4-year schools vs admission rate', xlab='Fra
ction of students admitted', ylab='Fraction of students completing program within 6 years')
```



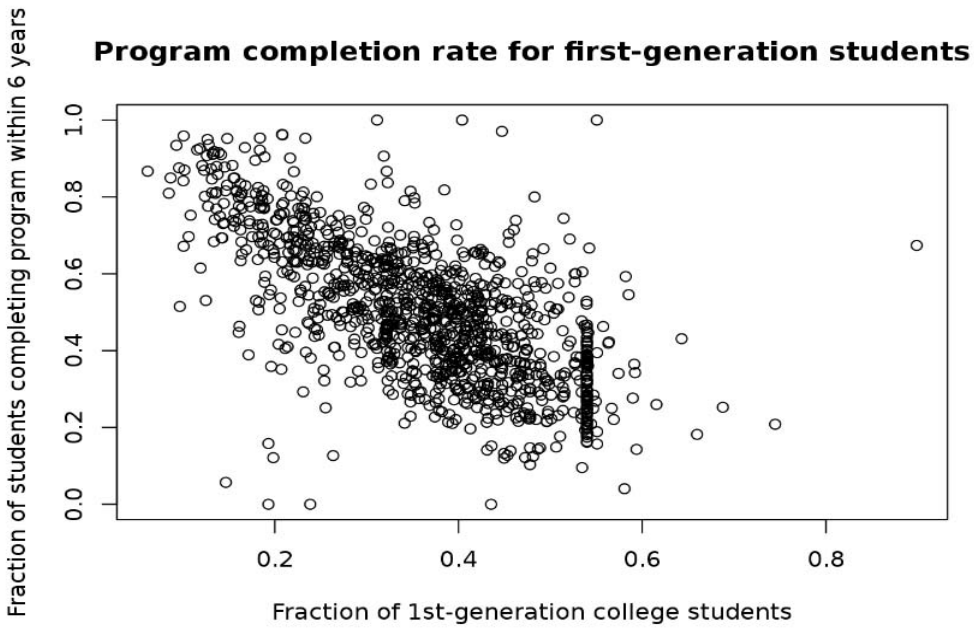
```
plot(train$COSTT4_A, train$C150_4, main='Program completion rate in 4-year schools vs cost of attendance', xlab='Cos
t of attendance/year', ylab='Fraction of students completing program within 6 years')
```




```
plot(train$INC_PCT_LO, train$C150_4, main='Program completion rate for students from low-income families', xlab='Fra
ction of students from families with income <$30000/yr', ylab='Fraction of students completing program within 6 year
s')
```



```
plot(train$PAR_ED_PCT_1STGEN, train$C150_4, main='Program completion rate for first-generation students', xlab='Frac
tion of 1st-generation college students', ylab='Fraction of students completing program within 6 years')
```



Prediction Models

We use four different machine learning models in our analysis for predicting the target variable (Completion rate for first-time, full-time students at four-year institutions (150% of expected time to completion/6 years) **C150_4**-VOI). We divide the entire dataset into training and test set. We train the model on training data and evaluate the model on the test data. We use confusion matrix as the metric for evaluating the performance of the model.

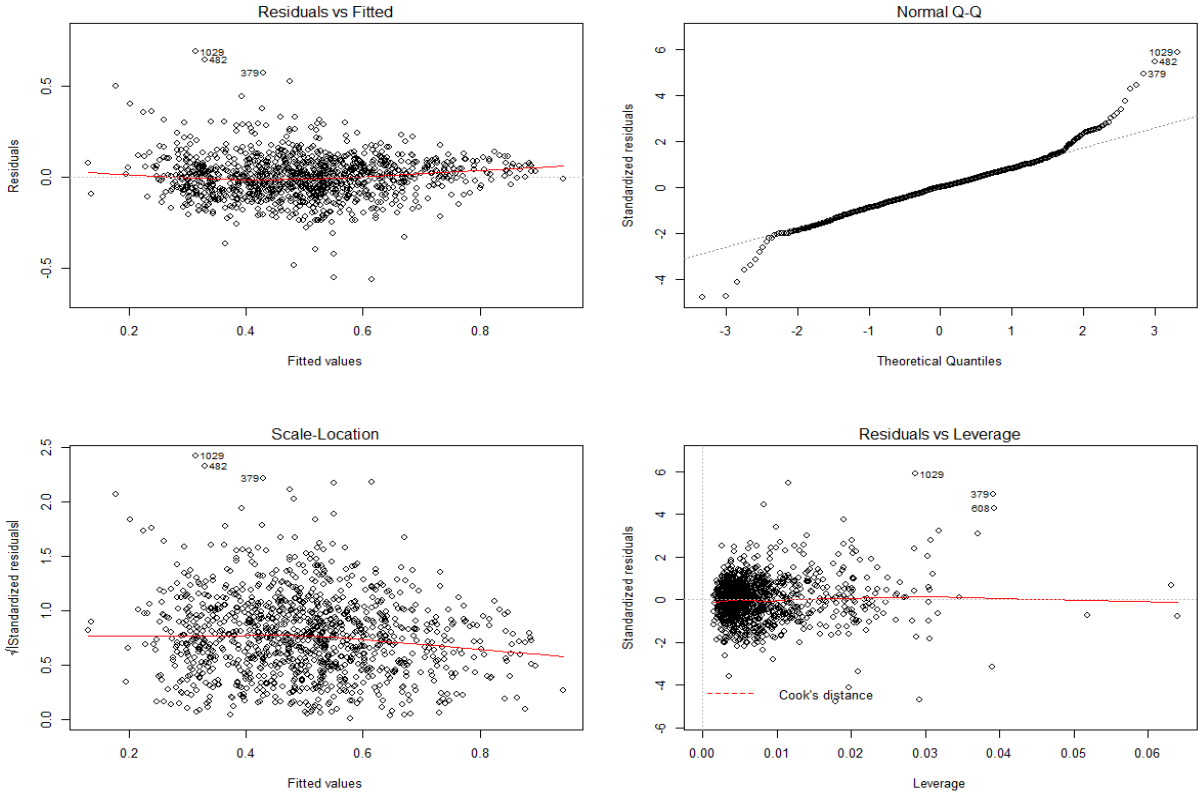
5.1 Multiple Linear Regression

In multiple linear regression we model the relationship between a single independent variable and a target variable. In this analysis, since we have multiple candidate features affecting the target variable, we use a multivariate linear regression model. The following is the snap of the multi linear regression model on my above dataset:

```
lmfit =lm(C150_4~.-Year,data=train )
summary(lmfit)
```

```
##
## Call:
## lm(formula = C150_4 ~ . - Year, data = train)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.55867 -0.06987  0.00198  0.06702  0.68567
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         8.284e-01  2.886e-02  28.702  < 2e-16 ***
## ADM_RATE_ALL       -1.543e-01  2.138e-02  -7.220 9.50e-13 ***
## COSTT4_A            4.357e-06  4.060e-07  10.732  < 2e-16 ***
## UGDS_BLACK         -7.142e-02  2.235e-02  -3.195  0.00143 **
## PPTUG_EF           -1.983e-01  3.662e-02  -5.415 7.48e-08 ***
## INC_PCT_LO         -3.698e-01  3.908e-02  -9.461  < 2e-16 ***
## UG25abv             1.672e-01  3.435e-02   4.869 1.28e-06 ***
## PAR_ED_PCT_1STGEN -3.691e-01  5.280e-02  -6.991 4.65e-12 ***
## PCTFLOAN           -1.082e-01  2.315e-02  -4.675 3.29e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1182 on 1134 degrees of freedom
## Multiple R-squared:  0.5816, Adjusted R-squared:  0.5786
## F-statistic:   197 on 8 and 1134 DF,  p-value: < 2.2e-16
```

From the above output, we can see that all the variables involved in the model are significant (p-value<0.05). Also the R-Square of the model is 58.16% which is quite good and can be used for predicting the completion rate in this dataset. As for the coefficients, we can see that most of them are negative. The plots of this multi-regression model are below:



From the above plots we can see a quite good linear relationship between residuals and fitted values used in this multi-linear regression model. Also the quantile-quantile plots depicts the normality of this model. Even the outliers are shown in the respective plots.

Single decision tree

I train a single decision tree using the **'rpart'** library with default training parameters. The most discriminating variables are seen to be:

[1] fraction of students from low-income families (INC_PCT_LO)
[2] total cost of attendance per year (COSTT4_A)
[3] fraction of first-generation college goers in the family (PAR_ED_PCT_1STGEN)
[4] rate of admission (ADM_RATE_ALL)

I evaluate the performance on the 2013 data. I define a correct prediction as the predicted completion rate for a given college being within 10% of the actual rate in 2013. The performance of a single tree is not very good: **64% of predictions are correct**.

```
fol = formula(C150_4 ~ ADM_RATE_ALL + COSTT4_A + UGDS_BLACK + PPTUG_EF + INC_PCT_LO + UG25abv + PAR_ED_PCT_1STGEN +
PCTFLOAN)

model_tree = rpart(fol, method="anova", data=train)

test <- dbGetQuery(db, "
SELECT Year, ADM_RATE_ALL, COSTT4_A, UGDS_BLACK, PPTUG_EF, INC_PCT_LO, UG25abv, PAR_ED_PCT_1STGEN, PCTFLOAN, C150_4
FROM Scorecard
WHERE Year = 2013
    AND COSTT4_A != 'PrivacySuppressed' AND COSTT4_A IS NOT NULL
    AND ADM_RATE_ALL != 'PrivacySuppressed' AND ADM_RATE_ALL IS NOT NULL
    AND UGDS_BLACK != 'PrivacySuppressed' AND UGDS_BLACK IS NOT NULL
    AND PPTUG_EF != 'PrivacySuppressed' AND PPTUG_EF IS NOT NULL
    AND INC_PCT_LO != 'PrivacySuppressed' AND INC_PCT_LO IS NOT NULL
    AND UG25abv != 'PrivacySuppressed' AND UG25abv IS NOT NULL
    AND PAR_ED_PCT_1STGEN != 'PrivacySuppressed' AND PAR_ED_PCT_1STGEN IS NOT NULL
    AND PCTFLOAN != 'PrivacySuppressed' AND PCTFLOAN IS NOT NULL
    AND C150_4 != 'PrivacySuppressed' AND C150_4 IS NOT NULL
")

pred_tree = predict(model_tree, newdata=test)
accu = abs(pred_tree - test$C150_4) < 0.1
frac = sum(accu)/length(accu)
print(frac)
```

```
## [1] 0.6423077
```

Random Forest

I then train a random forest of trees, again using default parameters. The four highest ranked variables, using the **importance()** function, are:

[1] fraction of students from low-income families (INC_PCT_LO)
[2] fraction of first-generation college goers in the family (PAR_ED_PCT_1STGEN)

[3] total cost of attendance per year (COSTT4_A)
[4] fraction of students above 25 years of age (UG25abv)

The performance is better: **76% of predictions are correct**.

```
model_forest = randomForest(fol, data=train)
pred_forest = predict(model_forest, newdata=test)
accu = abs(pred_forest - test$C150_4) < 0.1
frac = sum(accu)/length(accu)
print(frac)
```

```
## [1] 0.7615385
```

SVM (Support Vector Machines)

Finally, I train an SVM using the 'e1071' library using default parameter values. The performance is comparable with that of the random forest: **74% of predictions are correct**. But a tuning of parameters would yield better results with this method.

```
model_svm = svm(fol, data=train)
pred_svm = predict(model_svm, newdata=test)
accu = abs(pred_svm - test$C150_4) < 0.1
frac = sum(accu)/length(accu)
print(frac)
```

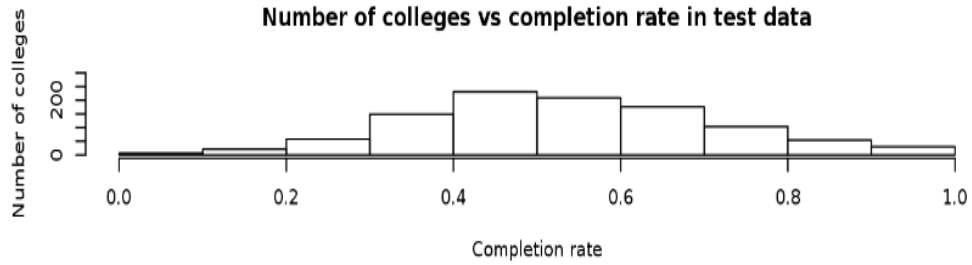
```
## [1] 0.7442308
```

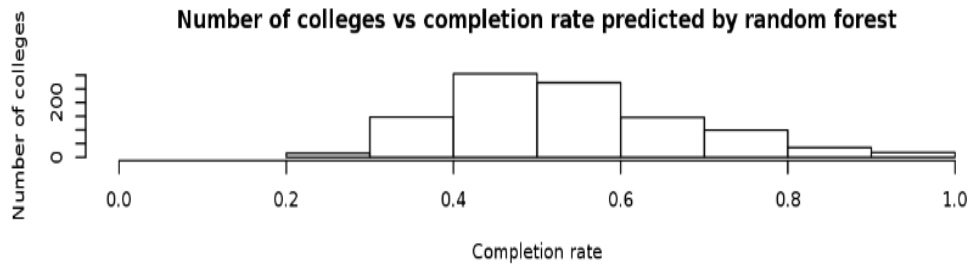
The number of colleges as a function of the completion rate are shown from the test data, from the random forest prediction and from the SVM prediction. Both methods perform badly for large values of the completion rate.

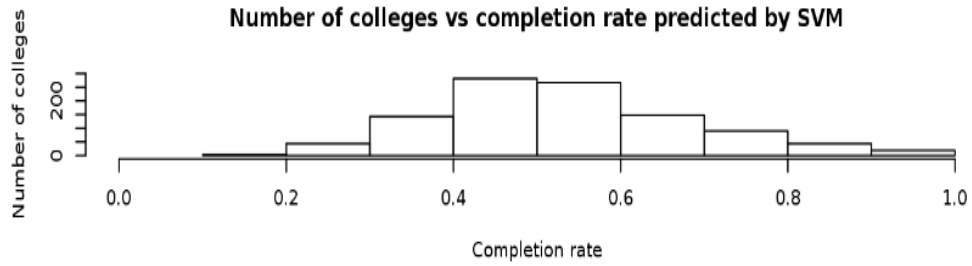
The next step in the analysis is to look at distributions of the training variables separately for low and high completion rates, and try to understand why the training is poorer in the latter case.

```
par(mfrow=c(3,1))
hist(test$C150_4, ylim=c(0,300), main='Number of colleges vs completion rate in test data', xlab='Completion rate',
ylab='Number of colleges')
hist(pred_forest, xlim=c(0,1), ylim=c(0,300), breaks=10, main='Number of colleges vs completion rate predicted by ra
ndom forest', xlab='Completion rate', ylab='Number of colleges')
hist(pred_svm, xlim=c(0,1), ylim=c(0,300), breaks=10, main='Number of colleges vs completion rate predicted by SVM',
 xlab='Completion rate', ylab='Number of colleges')
```

The below are the histogram for the above code which help us analyse our theory about the training data:







Conclusion

I used College Scorecard dataset to predict the fraction of students in 4-year colleges who complete the program in less than 6 years. The dataset had large number of missing values so we did pre-processing and reduced the number of features and colleges. I found the most significant features for predicting the fraction of students in 4-year colleges using multiple hypothesis testing and correlation tests. For prediction, we used Multiple-linear regression, Decision Tress, Random Forests and Support Vector Machines methods.

As I commented above, the data are not very reliable owing to the large number of privacy suppressions. To get an idea of the spread of the data (statistical and systematic components convolved), I ran my test all over again using the above mentioned methods on 2009 data. The fraction of correct predictions are thus:

Random forest: 65%

SVM: 69%

Which are within ~10% of the 2013 numbers.

The important conclusions of this study are not the performance of the MVAs, but the factors that they show to affect program completion rates the most. In general, they agree with our naive expectations as to why someone may fail to complete a college program.

Limitations and Ethical Implications

The main limitation of the Scorecard data stems from its focus on the earnings outcomes of federal aid recipients rather than of all students. Despite this limitation, however, the Scorecard provides consistent information across universities on critical measures regarding access, affordability and outcomes.

Based on the data dictionary, I started with 14 features in my dataset to assess the completion rate of the students. I collected these variables and examine them for availability, integrity and correlations. Some of the variables were privacy-suppressed for the majority of cases, yielding <500 records. Some of them then had a very small correlation with the VOI, so that this is excluded as well. That left me with 8 training variables. Given more data and the accuracy of the same would have really helped in consistent results throughout the years which in-turn would have helped to build a more robust model without outliers and normally distributed.

For many elements, based on our varied results it seems data is not stable from year to year and representative of a certain number of students. For most elements, data are pooled across years to reduce year-over-year variability in figures (i.e. repayment rate, completion rate, earnings). Moreover, for elements that we expect to publish for consumer use, a separate version of the element is available that suppresses the data for schools with fewer than 30 students in the denominator. All National Student Loan Data System (NSLDS) and Treasury elements are protected for privacy purposes; any data not reported in order to protect an individual's privacy are shown as PrivacySuppressed.

Additionally, many elements are available only for Title IV recipients, or students who receive federal grants and loans. While these data are reported at the individual level to NSLDS and used to distribute federal aid, they are published only at the aggregate institutional level. While some schools report these data at the campus level (C150_4-VOI) aggregating those I this data has not been done and have been placed in another dataset.

References

[1] https://collegescorecard.ed.gov/data/