Machine Learning Engineer Nanodegree

Capstone Proposal

Sanjeev Yadav

20 February 2018

Proposal

Domain Background

Supervised learning is one of the most popular areas of machine learning in which much development has already taken place. In this project I am trying to identify the university-level factors which predict the presence of a strong retention and graduation rate. As the leader of the big data revolution, Google gathers information through clicks on the Internet and uses this information to personalize advertising to individual users[1]. Academia will use the same model in the learning process to customize courses right down to the level of the individual. Some companies, such as the nonprofit testing firm ETS, are already harnessing data to develop predetermined learning trees to track certain responses to questions that imply mastery of specific aspects of material, thus allowing educators to organize assignments based on those answers[1]. In the United States, universities and colleges face tremendous pressures in terms of their business models, the mobility of students, the growing disillusionment with four-year degrees and the cost of higher education[1]. By paying specific attention to important factors, a university can increase its education status.

The link to my data source is here. The name of the file is data.csv.

Data was collected from data.gov, but for easy navigation we have pushed it to the GitHub repo.

Problem Statement

One of the most pressing issues facing American universities is the number of students who fail to graduate. Nearly one out of five four-year institutions graduate fewer than one-third of its first-time, full-time degree-seeking first-year students within six years. Although there are various explanations for attrition, I will try to identify the most important feature which affects the retention and graduation rates in 4-year institutions.

We have two target variables:

1. Graduation rate, and
2. Retention rate

Both are continuous variable so this is a regression task. We will train same regression models for both target variables but the final model will be chosen based on the `r2_score`. It may be the case that one model works good for graduation rate and some other works good for retention rate.

Datasets and Inputs

The dataset contains 2 csv files. They are:

- data.csv

It contains the input data with 123 variables and 7593 observations. I will create training set and testing set from this data after performing data preprocessing.

Features:

UNITID: Unit ID for institution

OPEID: 8-digit OPEID ID for institution

OPEID6: 6-digit OPEID for institution

INSTNM: Institution name

CITY: city

STABBR: State postcode

INSTURL: URL for instution's homepage

NPCURL: URL for institution's net price calculator

HCM2: Schools that are on Heightened Cash Monitoring 2 by the Department of Education

PREDDEG: Predominant undergraduate degree awarded. Can take 5 values:

1. Not classified
2. Predominantly certificate-degree granting
3. Predominantly associate's-degree granting
4. Predominantly bachelor's-degree granting
5. Entirely graduate-degree granting

HIGHDEG: Highest degree awarded. Can take 5 values:

1. Non-degree-granting
2. Certificate degree
3. Associate degree
4. Bachelor's degree
5. Graduate degree

CONTROL: Control of institution. Can take 3 values:

1. Public
2. Private non-profit
3. Private for-profit

LOCALE: Locale of institution. Can take 12 values:

1. City: Large (population of 250,000 or more)
2. City: Midsize (population of at least 100,000 but less than 250,000)
3. City: Small (population less than 100,000)
4. Suburb: Large (outside principal city, in urbanized area with population of 250,000 or more)
5. Suburb: Midsize (outside principal city, in urbanized area with population of at least 100,000 but less than 250,000)
6. Suburb: Small (outside principal city, in urbanized area with population less than 100,000)
7. Town: Fringe (in urban cluster up to 10 miles from an urbanized area)
8. Town: Distant (in urban cluster more than 10 miles and up to 35 miles from an urbanized area)
9. Town: Remote (in urban cluster more than 35 miles from an urbanized area)
10. Rural: Fringe (rural territory up to 5 miles from an urbanized area or up to 2.5 miles from an urban cluster)
11. Rural: Distant (rural territory more than 5 miles but up to 25 miles from an urbanized area or more than 2.5 and up to 10 miles from an urban cluster)
12. Rural: Remote (rural territory more than 25 miles from an urbanized area and more than 10 miles from an urban cluster)

HBCU: Flag for historically Black College and University.

PBI: Flag for predominantly black institution.

ANNHI: Flag for Alaska Native Native Hawaiian serving institution.

TRIBAL: Flag for tribal college and university

AANAPII: Flag for Asian American Native American Pacific Islander-serving institution

HSI: Flag for Hispanic-serving institution

NANTI: Flag for Native American non-tribal institution

MENONLY: Flag for men-only college

WOMENONLY: Flag for women-only college

RELAFFIL: Religious affiliation of the institution. It can take 65 values:

1. Not reported
2. Not applicable
3. American Evangelical Lutheran Church
4. African Methodist Episcopal Zion Church
5. Assemblies of God Church
6. Brethren Church
7. Roman Catholic
8. Wisconsin Evangelical Lutheran Synod
9. Christ and Missionary Alliance Church
10. Christian Reformed Church
11. Evangelical Congregational Church
12. Evangelical Covenant Church of America

13. Evangelical Free Church of America
14. Evangelical Lutheran Church
15. International United Pentecostal Church
16. Free Will Baptist Church
17. Interdenominational
18. Mennonite Brethren Church
19. Moravian Church
20. North American Baptist
21. Pentecostal Holiness Church
22. Christian Churches and Churches of Christ
23. Reformed Church in America
24. Episcopal Church, Reformed
25. African Methodist Episcopal
26. American Baptist
27. American Lutheran
28. Baptist
29. Christian Methodist Episcopal
30. Church of God
31. Church of Brethren
32. Church of the Nazarene
33. Cumberland Presbyterian
34. Christian Church (Disciples of Christ)
35. Free Methodist
36. Friends
37. Presbyterian Church (USA)
38. Lutheran Church in America
39. Lutheran Church - Missouri Synod
40. Mennonite Church
41. United Methodist
42. Protestant Episcopal
43. Churches of Christ
44. Southern Baptist
45. United Church of Christ
46. Protestant, not specified
47. Multiple Protestant Denomination
48. Other Protestant
49. Jewish
50. Reformed Presbyterian Church
51. United Brethren Church
52. Missionary Church Inc
53. Undenominational
54. Wesleyan
55. Greek Orthodox
56. Russian Orthodox
57. Unitarian Universalist
58. Latter Day Saints (Mormon Church)
59. Seventh Day Adventists
60. The Presbyterian Church in America
61. Other (none of the above)
62. Original Free Will Baptist
63. Ecumenical Christian

64. Evangelical Christian
65. Presbyterian

SATVR25: 25th percentile of SAT scores at the institution (critical reading)

SATVR75: 75th percentile of SAT scores at the institution (critical reading)

SATMT25: 25th percentile of SAT scores at the institution (math)

SATMT75: 75th percentile of SAT scores at the institution (math)

SATWR25: 25th percentile of SAT scores at the institution (writing)

SATWR75: 75th percentile of SAT scores at the institution (writing)

SATVRMID: Midpoint of SAT scores at the institution (critical reading)

SATMTMID: Midpoint of SAT scores at the institution (math)

SATWRMID: Midpoint of SAT scores at the institution (writing)

ACTCM25: 25th percentile of the ACT cumulative score

ACTCM75: 75th percentile of the ACT cumulative score

ACTEN25: 25th percentile of the ACT English score

ACTEN75: 75th percentile of the ACT English score

ACTMT25: 25th percentile of the ACT math score

ACTMT75: 75th percentile of the ACT math score

ACTWR25: 25th percentile of the ACT writing score

ACTWR75: 75th percentile of the ACT writing score

ACTCMMID: Midpoint of the ACT cumulative score

ACTENMID: Midpoint of the ACT English score

ACTMTMID: Midpoint of the ACT math score

ACTWRMID: Midpoint of the ACT writing score

SAT_AVG: Average SAT equivalent score of students admitted

SAT_AVG_ALL:Average SAT equivalent score of students admitted for all campuses rolled up to the 6-digit OPE ID

PCIP01: Percentage of degrees awarded in Agriculture, Agriculture Operations, And Related Sciences.

PCIP03: Percentage of degrees awarded in Natural Resources And Conservation.

PCIP04: Percentage of degrees awarded in Architecture And Related Services.

PCIP05: Percentage of degrees awarded in Area, Ethnic, Cultural, Gender, And Group Studies.

PCIP09: Percentage of degrees awarded in Communication, Journalism, And Related Programs.

PCIP10: Percentage of degrees awarded in Communications Technologies/Technicians And Support Services.

PCIP11: Percentage of degrees awarded in Computer And Information Sciences And Support Services.

PCIP12: Percentage of degrees awarded in Personal And Culinary Services.

PCIP13: Percentage of degrees awarded in Education.

PCIP14: Percentage of degrees awarded in Engineering.

PCIP15: Percentage of degrees awarded in Engineering Technologies And Engineering-Related Fields.

PCIP16: Percentage of degrees awarded in Foreign Languages, Literatures, And Linguistics.

PCIP19: Percentage of degrees awarded in Family And Consumer Sciences/Human Sciences.

PCIP22: Percentage of degrees awarded in Legal Professions And Studies.

PCIP23: Percentage of degrees awarded in English Language And Literature/Letters.

PCIP24: Percentage of degrees awarded in Liberal Arts And Sciences, General Studies And Humanities.

PCIP25: Percentage of degrees awarded in Library Science.

PCIP26: Percentage of degrees awarded in Biological And Biomedical Sciences.

PCIP27: Percentage of degrees awarded in Mathematics And Statistics.

PCIP29: Percentage of degrees awarded in Military Technologies And Applied Sciences.

PCIP30: Percentage of degrees awarded in Multi/Interdisciplinary Studies.

PCIP31: Percentage of degrees awarded in Parks, Recreation, Leisure, And Fitness Studies.

PCIP38: Percentage of degrees awarded in Philosophy And Religious Studies.

PCIP39: Percentage of degrees awarded in Theology And Religious Vocations.

PCIP40: Percentage of degrees awarded in Physical Sciences.

PCIP41: Percentage of degrees awarded in Science Technologies/Technicians.

PCIP42: Percentage of degrees awarded in Psychology.

PCIP43: Percentage of degrees awarded in Homeland Security, Law Enforcement, Firefighting And Related Protective Services.

PCIP44: Percentage of degrees awarded in Public Administration And Social Service Professions.

PCIP45: Percentage of degrees awarded in Social Sciences.

PCIP46: Percentage of degrees awarded in Construction Trades.

PCIP47: Percentage of degrees awarded in Mechanic And Repair Technologies/Technicians.

PCIP48: Percentage of degrees awarded in Precision Production.

PCIP49: Percentage of degrees awarded in Transportation And Materials Moving.

PCIP50: Percentage of degrees awarded in Visual And Performing Arts.

PCIP51: Percentage of degrees awarded in Health Professions And Related Programs.

PCIP52: Percentage of degrees awarded in Business, Management, Marketing, And Related Support Services.

PCIP54: Percentage of degrees awarded in History.

DISTANCEONLY: Flag for distance-education-only education

UGDS: Enrollment of undergraduate certificate/degree-seeking students

UGDS_WHITE: Total share of enrollment of undergraduate degree-seeking students who are white

UGDS_BLACK: Total share of enrollment of undergraduate degree-seeking students who are black

UGDS_HISP: Total share of enrollment of undergraduate degree-seeking students who are Hispanic

UGDS_ASIAN: Total share of enrollment of undergraduate degree-seeking students who are Asian

UGDS_AIAN: Total share of enrollment of undergraduate degree-seeking students who are American Indian/Alaska Native

UGDS_NHPI: Total share of enrollment of undergraduate degree-seeking students who are Native

Hawaiian/Pacific Islander

UGDS_2MOR: Total share of enrollment of undergraduate degree-seeking students who are two or more races

UGDS_NRA: Total share of enrollment of undergraduate degree-seeking students who are non-resident aliens

UGDS_UNKN: Total share of enrollment of undergraduate degree-seeking students whose race is unknown

PPTUG_EF: Share of undergraduate, degree-/certificate-seeking students who are part-time

CURROPER: Flag for currently operating institution, 0=closed, 1=operating

NPT4_PUB: Average net price for Title IV institutions (public institutions)

NPT4_PRIV: Average net price for Title IV institutions (private for-profit and nonprofit institutions)

NPT41_PUB: Average net price for \$0-$30,000 family income (public institutions)

NPT42_PUB: Average net price for \$30,001-$48,000 family income (public institutions)

NPT43_PUB: Average net price for \$48,001-$75,000 family income (public institutions)

NPT44_PUB: Average net price for \$75,001-$110,000 family income (public institutions)

NPT45_PUB: Average net price for \$110,000+ family income (public institutions)

NPT41_PRIV: Average net price for \$0-$30,000 family income (private for-profit and nonprofit institutions)

NPT42_PRIV: Average net price for \$30,001-$48,000 family income (private for-profit and nonprofit institutions)

NPT43_PRIV: Average net price for \$48,001-$75,000 family income (private for-profit and nonprofit institutions)

NPT44_PRIV: Average net price for \$75,001-$110,000 family income (private for-profit and nonprofit institutions)

NPT45_PRIV: Average net price for \$110,000+ family income (private for-profit and nonprofit institutions)

PCTPELL: Percentage of undergraduates who receive a Pell Grant

PCTFLOAN: Percent of all undergraduate students receiving a federal student loan

UG25ABV: Percentage of undergraduates aged 25 and above

MD_EARN_WNE_P10: Median earnings of students working and not enrolled 10 years after entry

GT_25K_P6: Share of students earning over $25,000/year (threshold earnings) 6 years after entry

GRAD_DEBT_MDN_SUPP: Median debt of completers, suppressed for n=30

GRAD_DEBT_MDN10YR_SUPP: Median debt of completers expressed in 10-year monthly payments, suppressed for n=30

RPY_3YR_RT_SUPP: 3-year repayment rate, suppressed for n=30

Target variables:

1. For graduation rates we have two variables in our data. Let us see the difference between those two:

1.1. **rate_suppressed.four_year**

Completion rate for first-time, full-time students at four-year institutions (150% of expected time to completion) , pooled in two-year rolling averages and suppressed for small n size.

1.2. **rate_suppressed.lt_four_year_150percent**

Completion rate for first-time, full-time students at less-than-four-year institutions (150% of expected time to completion) , pooled in two-year rolling averages and suppressed for small n size

We will be making predictions for 4-year institutions.

1. For retention rates we have four variables in our data. Let us see the difference between them:

2.1. **retention_rate.four_year.full_time**

First-time, full-time student retention rate at four-year institutions.

2.2. **retention_rate.lt_four_year.full_time**

First-time, full-time student retention rate at less-than-four-year institutions.

2.3. **retention_rate.four_year.part_time**

First-time, part-time student retention rate at four-year institutions

2.4. **retention_rate.lt_four_year.part_time**

First-time, part-time student retention rate at four-year institutions

Retention rate is for full-time students and we are making predictions for 4-year institutions. So, our target variable is **retention_rate.four_year.full_time**.

So there are 2 response variables:

1. **rate_suppressed.four_year**: This follows a normal distribution.
2. **retention_rate.four_year.full_time**: This follow a left skewed distribution.

- metadata.xlsx

This file contains information about the variables in our input data. It has information about extra variables also, so to improve readability we have highlighted the variables which are in our input data.

Solution Statement

We want to make predictions for two target variables: retention rates and graduation rates. We have around 100 features to choose from. First we will remove some irrelevant features. Our data set contains many null values. We will replace null values using median imputation method. Then we will perform feature scaling. Our target variables are continuous variables.

Then we will train a linear regressor on the training data for graduation rates and retention rates. We will get some features which have high importance. We will solidify our result by using other supervised regressors and picking the best out of them. Best here means the one with highest value of `r2_score`. Other supervised regressors which will be used are:

1. AdaBoost Regressor
2. Decision Tree Regressor
3. Extra Trees Regressor
4. Gradient Boosting Regressor
5. Random Forest Regressor

Benchmark Model

I will consider benchmark model as the initial linear regressor. We will get the `r2_score` from this model. Then we will use other regression models to improve our score.

Evaluation Metrics

We will use `r2_score` as the metric for performance of our model. `feature_importances_` method from sklearn will show the importance of each feature in predicting the target variables.

Project Design

First method will be using a linear regressor. Before that we will perform the preliminary steps of data preprocessing. We will see the meaning of each variable. Then we will perform feature selection. In this step we will remove the variables which we think are not important for our analysis. Then we will handle null values. There are 2 types of null values: NaNs and "PrivacySuppressed". We will first convert "PrivacySuppressed" to NaN and then perform median imputation method to fill null values. We have categorical variables also. For categorical variable with less than 10 levels, we will create dummy variables. Mainly there are 4 categorical variables. Other categorical variables are just flags which contain two values: 0 and 1.

When identifying the response variable, we have 6 columns to choose from. We will do our study for 4 year full time students. So, this will narrow down our choice to 2 variables: one for graduation rate and other for retention rate.

Now we have our features and targets, we will normalize the data using `StandardScaler`. Then we will perform train-test split with test size of 0.2.

Then we will use the supervised regressors to get the important features. `r2_score` will show the performance of our model. We will use these important features to see the practical importance of these features.

References

1. https://www.usnews.com/opinion/articles/2013/08/15/why-big-data-not-moocs-will-revolutionize-education